B5

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 December 2000 (07.12.2000)

PCT

(10) International Publication Number
WO 00/73509 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US00/15404

(22) International Filing Date: 31 May 2000 (31.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

| | | |
|---|---|---|
| 60/137,161 | 1 June 1999 (01.06.1999) | US |
| 60/137,109 | 2 June 1999 (02.06.1999) | US |
| 60/137,258 | 2 June 1999 (02.06.1999) | US |
| 60/137,260 | 2 June 1999 (02.06.1999) | US |
| 60/137,113 | 2 June 1999 (02.06.1999) | US |
| 60/137,259 | 2 June 1999 (02.06.1999) | US |
| 60/137,114 | 2 June 1999 (02.06.1999) | US |
| 60/137,173 | 2 June 1999 (02.06.1999) | US |
| 60/137,337 | 3 June 1999 (03.06.1999) | US |
| 60/137,417 | 3 June 1999 (03.06.1999) | US |
| 60/137,396 | 3 June 1999 (03.06.1999) | US |
| 60/137,411 | 3 June 1999 (03.06.1999) | US |
| 60/147,436 | 4 August 1999 (04.08.1999) | US |
| 60/147,377 | 4 August 1999 (04.08.1999) | US |
| 60/147,549 | 5 August 1999 (05.08.1999) | US |
| 60/147,527 | 5 August 1999 (05.08.1999) | US |
| 60/147,520 | 5 August 1999 (05.08.1999) | US |
| 60/147,536 | 5 August 1999 (05.08.1999) | US |
| 60/147,530 | 5 August 1999 (05.08.1999) | US |
| 60/147,500 | 5 August 1999 (05.08.1999) | US |
| 60/147,547 | 5 August 1999 (05.08.1999) | US |
| 60/147,824 | 5 August 1999 (05.08.1999) | US |
| 60/147,541 | 5 August 1999 (05.08.1999) | US |
| 60/147,542 | 5 August 1999 (05.08.1999) | US |

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

| | |
|---|---|
| US | 60,137,161 (CIP) |
| Filed on | 1 June 1999 (01.06.1999) |
| US | 60/137,109 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,258 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,260 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,113 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,259 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,114 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,173 (CIP) |
| Filed on | 2 June 1999 (02.06.1999) |
| US | 60/137,337 (CIP) |
| Filed on | 3 June 1999 (03.06.1999) |
| US | 60/137,417 (CIP) |
| Filed on | 3 June 1999 (03.06.1999) |
| US | 60/137,396 (CIP) |
| Filed on | 3 June 1999 (03.06.1999) |
| US | 60/137,411 (CIP) |
| Filed on | 3 June 1999 (03.06.1999) |
| US | 60/147,436 (CIP) |
| Filed on | 4 August 1999 (04.08.1999) |
| US | 60/147,377 (CIP) |
| Filed on | 4 August 1999 (04.08.1999) |
| US | 60/147,549 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,527 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,520 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,536 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,530 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,547 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,824 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,541 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,542 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |
| US | 60/147,500 (CIP) |
| Filed on | 5 August 1999 (05.08.1999) |

(71) Applicant (*for all designated States except US*): INCYTE GENOMICS, INC. [US/US]; 3160 Porter Drive, Palo Alto, CA 94304 (US).

*[Continued on next page]*

(54) Title: MOLECULES FOR DIAGNOSTICS AND THERAPEUTICS

(57) Abstract: The present invention provides purified human polynucleotides for diagnostics and therapeutics (dithp). Also encompassed are the polypeptides (DITHP) encoded by dithp. The invention also provides for the use of dithp, or complements, oligonucleotides, or fragments thereof in diagnostic assays. The invention further provides for vectors and host cells containing dithp for the expression of DITHP. The invention additionally provides for the use of isolated and purified DITHP to induce antibodies and to screen libraries of compounds and the use of anti-DITHP antibodies in diagnostic assays. Also provided are microarrays containing dithp and methods of use.

WO 00/73509 A2

(72) **Inventors; and**

(75) **Inventors/Applicants** *(for US only)*: **HODGSON, David, M.** [US/US]; 567 Addison Avenue, Palo Alto, CA 94301 (US). **LINCOLN, Stephen, E.** [US/US]; 725 Sapphire Street, Redwood City, CA 94061 (US). **RUSSO, Frank, D.** [US/US]; 1583 Courdillaeras Road, Redwood City, CA 94062 (US). **SPIRO, Peter, A.** [US/US]; 3875 Park Boulevard, Apt. B16, Palo Alto, CA 94306 (US). **BANVILLE, Steven, C.** [US/US]; 604 San Diego Avenue, Sunnyvale, CA 95086 (US). **BRATCHER, Shawn, R.** [US/US]; 550 Ortega Avenue #B321, Mountain View, CA 94040 (US). **DUFOUR, Gerard, E.** [US/US]; 5327 Greenridge Road, Castro Valley, CA 94552-2619 (US). **COHEN, Howard, J.** [US/US]; 3272 Cowper Street, Palo Alto, CA 94306-3004 (US). **ROSEN, Bruce, H.** [US/US]; 177 Hanna Way, Menlo Park, CA 94025 (US). **CHALUP, Michael, S.** [US/US]; 183 Acalanes Drive, Apt. 6, Sunnyvale, CA 94086 (US). **HILLMAN, Jennifer, L.** [US/US]; 230 Monroe Drive #12, Mountain View, CA 94040 (US). **JONES, Anissa, L.** [US/US]; 445 South 15th Street, San Jose, CA 95112 (US). **YU, Jimmy, Y.** [US/US]; 37330 Portico Terrace, Fremont, CA 94536-7901 (US). **GREENAWALT, Lila, B.** [US/US]; 1596 Ballantree Way, San Jose, CA 95118-2106 (US). **PANZER, Scott, R.** [US/US]; 965 East El Camino #621, Sunnyvale, CA 94087 (US). **ROSEBERRY, Ann, M.** [US/US]; 725 Sapphire Street, Redwood City, CA 94061 (US). **WRIGHT, Rachel, J.** [NZ/US]; 339 Anna Way, Mountain View, CA 94043 (US). **DANIELS, Susan, E.** [GB/US]; 136 Seale Avenue, Palo Alto, CA 94301 (US).

(74) **Agents: HAMLET-COX, Diana** et al.; Incyte Genomics, Inc., 3160 Porter Drive, Palo Alto, CA 94304 (US).

(81) **Designated States** *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *Without international search report and to be republished upon receipt of that report.*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

# MOLECULES FOR DIAGNOSTICS AND THERAPEUTICS

## TECHNICAL FIELD

The present invention relates to human molecules for diagnostics and therapeutics and to the use of these sequences in the diagnosis, study, prevention, and treatment of diseases associated with human molecules.

## BACKGROUND OF THE INVENTION

The human genome is comprised of thousands of genes, many encoding gene products that function in the maintenance and growth of the various cells and tissues in the body. Aberrant expression or mutations in these genes and their products is the cause of, or is associated with, a variety of human diseases such as cancer and other cell proliferative disorders, autoimmune/inflammatory disorders, infections, developmental disorders, endocrine disorders, metabolic disorders, neurological disorders, gastrointestinal disorders, transport disorders, and connective tissue disorders. The identification of these genes and their products is the basis of an ever-expanding effort to find markers for early detection of diseases, and targets for their prevention and treatment. Therefore, these genes and their products are useful as diagnostics and therapeutics. These genes may encode, for example, enzyme molecules, molecules associated with growth and development, biochemical pathway molecules, extracellular information transmission molecules, receptor molecules, intracellular signaling molecules, membrane transport molecules, protein modification and maintenance molecules, nucleic acid synthesis and modification molecules, adhesion molecules, antigen recognition molecules, secreted and extracellular matrix molecules, cytoskeletal molecules, ribosomal molecules, electron transfer associated molecules, transcription factor molecules, chromatin molecules, cell membrane molecules, and organelle associated molecules.

For example, cancer represents a type of cell proliferative disorder that affects nearly every tissue in the body. A wide variety of molecules, either aberrantly expressed or mutated, can be the cause of, or involved with, various cancers because tissue growth involves complex and ordered patterns of cell proliferation, cell differentiation, and apoptosis. Cell proliferation must be regulated to maintain both the number of cells and their spatial organization. This regulation depends upon the appropriate expression of proteins which control cell cycle progression in response to extracellular signals such as growth factors and other mitogens, and intracellular cues such as DNA damage or nutrient starvation. Molecules which directly or indirectly modulate cell cycle progression fall into several categories, including growth factors and their receptors, second messenger and signal transduction proteins, oncogene products, tumor-suppressor proteins, and mitosis-promoting factors. Aberrant expression or mutations in any of these gene products can result in cell proliferative disorders

such as cancer. Oncogenes are genes generally derived from normal genes that, through abnormal expression or mutation, can effect the transformation of a normal cell to a malignant one (oncogenesis). Oncoproteins, encoded by oncogenes, can affect cell proliferation in a variety of ways and include growth factors, growth factor receptors, intracellular signal transducers, nuclear transcription factors, and cell-cycle control proteins. In contrast, tumor-suppressor genes are involved in inhibiting cell proliferation. Mutations which cause reduced function or loss of function in tumor-suppressor genes result in aberrant cell proliferation and cancer. Although many different genes and their products have been found to be associated with cell proliferative disorders such as cancer, many more may exist that are yet to be discovered.

DNA-based arrays can provide a simple way to explore the expression of a single polymorphic gene or a large number of genes. When the expression of a single gene is explored, DNA-based arrays are employed to detect the expression of specific gene variants. For example, a p53 tumor suppressor gene array is used to determine whether individuals are carrying mutations that predispose them to cancer. A cytochrome p450 gene array is useful to determine whether individuals have one of a number of specific mutations that could result in increased drug metabolism, drug resistance or drug toxicity.

DNA-based array technology is especially relevant for the rapid screening of expression of a large number of genes. There is a growing awareness that gene expression is affected in a global fashion. A genetic predisposition, disease or therapeutic treatment may affect, directly or indirectly, the expression of a large number of genes. In some cases the interactions may be expected, such as when the genes are part of the same signaling pathway. In other cases, such as when the genes participate in separate signaling pathways, the interactions may be totally unexpected. Therefore, DNA-based arrays can be used to investigate how genetic predisposition, disease, or therapeutic treatment affects the expression of a large number of genes.

**Enzyme Molecules**

SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, and SEQ ID NO:5 encode, for example, human enzyme molecules.

The cellular processes of biogenesis and biodegradation involve a number of key enzyme classes including oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. These enzyme classes are each comprised of numerous substrate-specific enzymes having precise and well regulated functions. These enzymes function by facilitating metabolic processes such as glycolysis, the tricarboxylic cycle, and fatty acid metabolism; synthesis or degradation of amino acids, steroids, phospholipids, alcohols, etc.; regulation of cell signalling, proliferation, inflamation, apoptosis, etc., and through catalyzing critical steps in DNA replication and repair, and the process of translation.

Oxidoreductases

Many pathways of biogenesis and biodegradation require oxidoreductase (dehydrogenase or reductase) activity, coupled to the reduction or oxidation of a donor or acceptor cofactor. Potential cofactors include cytochromes, oxygen, disulfide, iron-sulfur proteins, flavin adenine dinucleotide (FAD), and the nicotinamide adenine dinucleotides NAD and NADP (Newsholme, E.A. and A.R. Leech (1983) Biochemistry for the Medical Sciences, John Wiley and Sons, Chichester, U.K., pp. 779-793). Reductase activity catalyzes the transfer of electrons between substrate(s) and cofactor(s) with concurrent oxidation of the cofactor. The reverse dehydrogenase reaction catalyzes the reduction of a cofactor and consequent oxidation of the substrate. Oxidoreductase enzymes are a broad superfamily of proteins that catalyze numerous reactions in all cells of organisms ranging from bacteria to plants to humans. These reactions include metabolism of sugar, certain detoxification reactions in the liver, and the synthesis or degradation of fatty acids, amino acids, glucocorticoids, estrogens, androgens, and prostaglandins. Different family members are named according to the direction in which their reactions are typically catalyzed; thus they may be referred to as oxidoreductases, oxidases, reductases, or dehydrogenases. In addition, family members often have distinct cellular localizations, including the cytosol, the plasma membrane, mitochondrial inner or outer membrane, and peroxisomes.

Short-chain alcohol dehydrogenases (SCADs) are a family of dehydrogenases that only share 15% to 30% sequence identity, with similarity predominantly in the coenzyme binding domain and the substrate binding domain. In addition to the well-known role in detoxification of ethanol, SCADs are also involved in synthesis and degradation of fatty acids, steroids, and some prostaglandins, and are therefore implicated in a variety of disorders such as lipid storage disease, myopathy, SCAD deficiency, and certain genetic disorders. For example, retinol dehydrogenase is a SCAD-family member (Simon, A. et al. (1995) J. Biol. Chem. 270:1107-1112) that converts retinol to retinal, the precursor of retinoic acid. Retinoic acid, a regulator of differentiation and apoptosis, has been shown to down-regulate genes involved in cell proliferation and inflammation (Chai, X. et al. (1995) J. Biol. Chem. 270:3900-3904). In addition, retinol dehydrogenase has been linked to hereditary eye diseases such as autosomal recessive childhood-onset severe retinal dystrophy (Simon, A. et al. (1996) Genomics 36:424-430).

Propagation of nerve impulses, modulation of cell proliferation and differentiation, induction of the immune response, and tissue homeostasis involve neurotransmitter metabolism (Weiss, B. (1991) Neurotoxicology 12:379-386; Collins, S.M. et al. (1992) Ann. N.Y. Acad. Sci. 664:415-424; Brown, J.K. and H. Imam (1991) J. Inherit. Metab. Dis. 14:436-458). Many pathways of neurotransmitter metabolism require oxidoreductase activity, coupled to reduction or oxidation of a cofactor, such as $NAD^+$/NADH (Newsholme, E.A. and A.R. Leech (1983) Biochemistry for the Medical Sciences, John Wiley and Sons, Chichester, U.K. pp. 779-793). Degradation of

catecholamines (epinephrine or norepinephrine) requires alcohol dehydrogenase (in the brain) or aldehyde dehydrogenase (in peripheral tissue). $NAD^+$-dependent aldehyde dehydrogenase oxidizes 5-hydroxyindole-3-acetate (the product of 5-hydroxytryptamine (serotonin) metabolism) in the brain, blood platelets, liver and pulmonary endothelium (Newsholme, supra, p. 786). Other neurotransmitter degradation pathways that utilize $NAD^+$/NADH-dependent oxidoreductase activity include those of L-DOPA (precursor of dopamine, a neuronal excitatory compound), glycine (an inhibitory neurotransmitter in the brain and spinal cord), histamine (liberated from mast cells during the inflammatory response), and taurine (an inhibitory neurotransmitter of the brain stem, spinal cord and retina) (Newsholme, supra, pp. 790, 792). Epigenetic or genetic defects in neurotransmitter metabolic pathways can result in a spectrum of disease states in different tissues including Parkinson disease and inherited myoclonus (McCance, K.L. and S.E. Huether (1994) Pathophysiology, Mosby-Year Book, Inc., St. Louis MO, pp. 402-404; Gundlach, A.L. (1990) FASEB J. 4:2761-2766).

Tetrahydrofolate is a derivatized glutamate molecule that acts as a carrier, providing activated one-carbon units to a wide variety of biosynthetic reactions, including synthesis of purines, pyrimidines, and the amino acid methionine. Tetrahydrofolate is generated by the activity of a holoenzyme complex called tetrahydrofolate synthase, which includes three enzyme activities: tetrahydrofolate dehydrogenase, tetrahydrofolate cyclohydrolase, and tetrahydrofolate synthetase. Thus, tetrahydrofolate dehydrogenase plays an important role in generating building blocks for nucleic and amino acids, crucial to proliferating cells.

3-Hydroxyacyl-CoA dehydrogenase (3HACD) is involved in fatty acid metabolism. It catalyzes the reduction of 3-hydroxyacyl-CoA to 3-oxoacyl-CoA, with concomitant oxidation of NAD to NADH, in the mitochondria and peroxisomes of eukaryotic cells. In peroxisomes, 3HACD and enoyl-CoA hydratase form an enzyme complex called bifunctional enzyme, defects in which are associated with peroxisomal bifunctional enzyme deficiency. This interruption in fatty acid metabolism produces accumulation of very-long chain fatty acids, disrupting development of the brain, bone, and adrenal glands. Infants born with this deficiency typically die within 6 months (Watkins, P. et al. (1989) J. Clin. Invest. 83:771-777; Online Mendelian Inheritance in Man (OMIM), #261515). The neurodegeneration that is characteristic of Alzheimer's disease involves development of extracellular plaques in certain brain regions. A major protein component of these plaques is the peptide amyloid-β (Aβ), which is one of several cleavage products of amyloid precursor protein (APP). 3HACD has been shown to bind the Aβ peptide, and is overexpressed in neurons affected in Alzheimer's disease. In addition, an antibody against 3HACD can block the toxic effects of Aβ in a cell culture model of Alzheimer's disease (Yan, S. et al. (1997) Nature 389:689-695; OMIM, #602057).

Steroids, such as estrogen, testosterone, corticosterone, and others, are generated from a

common precursor, cholesterol, and are interconverted into one another. A wide variety of enzymes act upon cholesterol, including a number of dehydrogenases. Steroid dehydrogenases, such as the hydroxysteroid dehydrogenases, are involved in hypertension, fertility, and cancer (Duax, W.L. and D. Ghosh (1997) Steroids 62:95-100). One such dehydrogenase is 3-oxo-5-α-steroid dehydrogenase (OASD), a microsomal membrane protein highly expressed in prostate and other androgen-responsive tissues. OASD catalyzes the conversion of testosterone into dihydrotestosterone, which is the most potent androgen. Dihydrotestosterone is essential for the formation of the male phenotype during embryogenesis, as well as for proper androgen-mediated growth of tissues such as the prostate and male genitalia. A defect in OASD that prevents the conversion of testosterone into dihydrotestosterone leads to a rare form of male pseudohermaphroditis, characterized by defective formation of the external genitalia (Andersson, S. et al. (1991) Nature 354:159-161; Labrie, F. et al. (1992) Endocrinology 131:1571-1573; OMIM #264600). Thus, OASD plays a central role in sexual differentiation and androgen physiology.

17β-hydroxysteroid dehydrogenase (17βHSD6) plays an important role in the regulation of the male reproductive hormone, dihydrotestosterone (DHTT). 17βHSD6 acts to reduce levels of DHTT by oxidizing a precursor of DHTT, 3α-diol, to androsterone which is readily glucuronidated and removed from tissues. 17βHSD6 is active with both androgen and estrogen substrates when expressed in embryonic kidney 293 cells. At least five other isozymes of 17βHSD have been identified that catalyze oxidation and/or reduction reactions in various tissues with preferences for different steroid substrates (Biswas, M.G. and D.W. Russell (1997) J. Biol. Chem. 272:15959-15966). For example, 17βHSD1 preferentially reduces estradiol and is abundant in the ovary and placenta. 17βHSD2 catalyzes oxidation of androgens and is present in the endometrium and placenta. 17βHSD3 is exclusively a reductive enzyme in the testis (Geissler, W.M. et al. (1994) Nat. Genet. 7:34-39). An excess of androgens such as DHTT can contribute to certain disease states such as benign prostatic hyperplasia and prostate cancer.

Oxidoreductases are components of the fatty acid metabolism pathways in mitochondria and peroxisomes. The main beta-oxidation pathway degrades both saturated and unsaturated fatty acids, while the auxiliary pathway performs additional steps required for the degradation of unsaturated fatty acids. The auxiliary beta-oxidation enzyme 2,4-dienoyl-CoA reductase catalyzes the removal of even-numbered double bonds from unsaturated fatty acids prior to their entry into the main beta-oxidation pathway. The enzyme may also remove odd-numbered double bonds from unsaturated fatty acids (Koivuranta, K.T. et al. (1994) Biochem. J. 304:787-792; Smeland, T.E. et al. (1992) Proc. Natl. Acad. Sci. USA 89:6673-6677). 2,4-dienoyl-CoA reductase is located in both mitochondria and peroxisomes. Inherited deficiencies in mitochondrial and peroxisomal beta-oxidation enzymes are associated with severe diseases, some of which manifest themselves soon after birth and lead to death

within a few years. Defects in beta-oxidation are associated with Reye's syndrome, Zellweger syndrome, neonatal adrenoleukodystrophy, infantile Refsum's disease, acyl-CoA oxidase deficiency, and bifunctional protein deficiency (Suzuki, Y. et al. (1994) Am. J. Hum. Genet. 54:36-43; Hoefler, supra; Cotran, R.S. et al. (1994) Robbins Pathologic Basis of Disease, W.B. Saunders Co., Philadelphia PA, p.866). Peroxisomal beta-oxidation is impaired in cancerous tissue. Although neoplastic human breast epithelial cells have the same number of peroxisomes as do normal cells, fatty acyl-CoA oxidase activity is lower than in control tissue (el Bouhtoury, F. et al. (1992) J. Pathol. 166:27-35). Human colon carcinomas have fewer peroxisomes than normal colon tissue and have lower fatty-acyl-CoA oxidase and bifunctional enzyme (including enoyl-CoA hydratase) activities than normal tissue (Cable, S. et al. (1992) Virchows Arch. B Cell Pathol. Incl. Mol. Pathol. 62:221-226). Another important oxidoreductase is isocitrate dehydrogenase, which catalyzes the conversion of isocitrate to a-ketoglutarate, a substrate of the citric acid cycle. Isocitrate dehydrogenase can be either NAD or NADP dependent, and is found in the cytosol, mitochondria, and peroxisomes. Activity of isocitrate dehydrogenase is regulated developmentally, and by hormones, neurotransmitters, and growth factors.

Hydroxypyruvate reductase (HPR), a peroxisomal 2-hydroxyacid dehydrogenase in the glycolate pathway, catalyzes the conversion of hydroxypyruvate to glycerate with the oxidation of both NADH and NADPH. The reverse dehydrogenase reaction reduces $NAD^+$ and $NADP^+$. HPR recycles nucleotides and bases back into pathways leading to the synthesis of ATP and GTP. ATP and GTP are used to produce DNA and RNA and to control various aspects of signal transduction and energy metabolism. Inhibitors of purine nucleotide biosynthesis have long been employed as antiproliferative agents to treat cancer and viral diseases. HPR also regulates biochemical synthesis of serine and cellular serine levels available for protein synthesis.

The mitochondrial electron transport (or respiratory) chain is a series of oxidoreductase-type enzyme complexes in the mitochondrial membrane that is responsible for the transport of electrons from NADH through a series of redox centers within these complexes to oxygen, and the coupling of this oxidation to the synthesis of ATP (oxidative phosphorylation). ATP then provides the primary source of energy for driving a cell's many energy-requiring reactions. The key complexes in the respiratory chain are NADH:ubiquinone oxidoreductase (complex I), succinate:ubiquinone oxidoreductase (complex II), cytochrome $c_1$-b oxidoreductase (complex III), cytochrome c oxidase (complex IV), and ATP synthase (complex V) (Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing, Inc., New York NY, pp. 677-678). All of these complexes are located on the inner matrix side of the mitochondrial membrane except complex II, which is on the cytosolic side. Complex II transports electrons generated in the citric acid cycle to the respiratory chain. The electrons generated by oxidation of succinate to fumarate in the citric acid cycle are transferred

through electron carriers in complex II to membrane bound ubiquinone (Q). Transcriptional regulation of these nuclear-encoded genes appears to be the predominant means for controlling the biogenesis of respiratory enzymes. Defects and altered expression of enzymes in the respiratory chain are associated with a variety of disease conditions.

Other dehydrogenase activities using NAD as a cofactor are also important in mitochondrial function. 3-hydroxyisobutyrate dehydrogenase (3HBD), important in valine catabolism, catalyzes the NAD-dependent oxidation of 3-hydroxyisobutyrate to methylmalonate semialdehyde within mitochondria. Elevated levels of 3-hydroxyisobutyrate have been reported in a number of disease states, including ketoacidosis, methylmalonic acidemia, and other disorders associated with deficiencies in methylmalonate semialdehyde dehydrogenase (Rougraff, P.M. et al. (1989) J. Biol. Chem. 264:5899-5903).

Another mitochondrial dehydrogenase important in amino acid metabolism is the enzyme isovaleryl-CoA-dehydrogenase (IVD). IVD is involved in leucine metabolism and catalyzes the oxidation of isovaleryl-CoA to 3-methylcrotonyl-CoA. Human IVD is a tetrameric flavoprotein that is encoded in the nucleus and synthesized in the cytosol as a 45 kDa precursor with a mitochondrial import signal sequence. A genetic deficiency, caused by a mutation in the gene encoding IVD, results in the condition known as isovaleric acidemia. This mutation results in inefficient mitochondrial import and processing of the IVD precursor (Vockley, J. et al. (1992) J. Biol. Chem. 267:2494-2501).

Transferases

Transferases are enzymes that catalyze the transfer of molecular groups. The reaction may involve an oxidation, reduction, or cleavage of covalent bonds, and is often specific to a substrate or to particular sites on a type of substrate. Transferases participate in reactions essential to such functions as synthesis and degradation of cell components, regulation of cell functions including cell signaling, cell proliferation, inflamation, apoptosis, secretion and excretion. Transferases are involved in key steps in disease processes involving these functions. Transferases are frequently classified according to the type of group transferred. For example, methyl transferases transfer one-carbon methyl groups, amino transferases transfer nitrogenous amino groups, and similarly denominated enzymes transfer aldehyde or ketone, acyl, glycosyl, alkyl or aryl, isoprenyl, saccharyl, phosphorous-containing, sulfur-containing, or selenium-containing groups, as well as small enzymatic groups such as Coenzyme A.

Acyl transferases include peroxisomal carnitine octanoyl transferase, which is involved in the fatty acid beta-oxidation pathway, and mitochondrial carnitine palmitoyl transferases, involved in fatty acid metabolism and transport. Choline O-acetyl transferase catalyzes the biosynthesis of the neurotransmitter acetylcholine.

Amino transferases play key roles in protein synthesis and degradation, and they contribute to

other processes as well. For example, the amino transferase 5-aminolevulinic acid synthase catalyzes the addition of succinyl-CoA to glycine, the first step in heme biosynthesis. Other amino transferases participate in pathways important for neurological function and metabolism. For example, glutamine-phenylpyruvate amino transferase, also known as glutamine transaminase K (GTK), catalyzes several reactions with a pyridoxal phosphate cofactor. GTK catalyzes the reversible conversion of L-glutamine and phenylpyruvate to 2-oxoglutaramate and L-phenylalanine. Other amino acid substrates for GTK include L-methionine, L-histidine, and L-tyrosine. GTK also catalyzes the conversion of kynurenine to kynurenic acid, a tryptophan metabolite that is an antagonist of the N-methyl-D-aspartate (NMDA) receptor in the brain and may exert a neuromodulatory function. Alteration of the kynurenine metabolic pathway may be associated with several neurological disorders. GTK also plays a role in the metabolism of halogenated xenobiotics conjugated to glutathione, leading to nephrotoxicity in rats and neurotoxicity in humans. GTK is expressed in kidney, liver, and brain. Both human and rat GTKs contain a putative pyridoxal phosphate binding site (ExPASy ENZYME: EC 2.6.1.64; Perry, S.J. et al. (1993) Mol. Pharmacol. 43:660-665; Perry, S. et al. (1995) FEBS Lett. 360:277-280; and Alberati-Giani, D. et al. (1995) J. Neurochem. 64:1448-1455). A second amino transferase associated with this pathway is kynurenine/α-aminoadipate amino transferase (AadAT). AadAT catalyzes the reversible conversion of α-aminoadipate and α-ketoglutarate to α-ketoadipate and L-glutamate during lysine metabolism. AadAT also catalyzes the transamination of kynurenine to kynurenic acid. A cytosolic AadAT is expressed in rat kidney, liver, and brain (Nakatani, Y. et al. (1970) Biochim. Biophys. Acta 198:219-228; Buchli, R. et al. (1995) J. Biol. Chem. 270:29330-29335).

Glycosyl transferases include the mammalian UDP-glucouronosyl transferases, a family of membrane-bound microsomal enzymes catalyzing the transfer of glucouronic acid to lipophilic substrates in reactions that play important roles in detoxification and excretion of drugs, carcinogens, and other foreign substances. Another mammalian glycosyl transferase, mammalian UDP-galactose-ceramide galactosyl transferase, catalyzes the transfer of galactose to ceramide in the synthesis of galactocerebrosides in myelin membranes of the nervous system. The UDP-glycosyl transferases share a conserved signature domain of about 50 amino acid residues (PROSITE: PDOC00359, http://expasy.hcuge.ch/sprot/prosite.html).

Methyl transferases are involved in a variety of pharmacologically important processes. Nicotinamide N-methyl transferase catalyzes the N-methylation of nicotinamides and other pyridines, an important step in the cellular handling of drugs and other foreign compounds. Phenylethanolamine N-methyl transferase catalyzes the conversion of noradrenalin to adrenalin. 6-O-methylguanine-DNA methyl transferase reverses DNA methylation, an important step in carcinogenesis. Uroporphyrin-III C-methyl transferase, which catalyzes the transfer of two methyl

groups from S-adenosyl-L-methionine to uroporphyrinogen III, is the first specific enzyme in the biosynthesis of cobalamin, a dietary enzyme whose uptake is deficient in pernicious anemia. Protein-arginine methyl transferases catalyze the posttranslational methylation of arginine residues in proteins, resulting in the mono- and dimethylation of arginine on the guanidino group. Substrates include histones, myelin basic protein, and heterogeneous nuclear ribonucleoproteins involved in mRNA processing, splicing, and transport. Protein-arginine methyl transferase interacts with proteins upregulated by mitogens, with proteins involved in chronic lymphocytic leukemia, and with interferon, suggesting an important role for methylation in cytokine receptor signaling (Lin, W.-J. et al. (1996) J. Biol. Chem. 271:15034-15044; Abramovich, C. et al. (1997) EMBO J. 16:260-266; and Scott, H.S. et al. (1998) Genomics 48:330-340).

Phosphotransferases catalyze the transfer of high-energy phosphate groups and are important in energy-requiring and -releasing reactions. The metabolic enzyme creatine kinase catalyzes the reversible phosphate transfer between creatine/creatine phosphate and ATP/ADP. Glycocyamine kinase catalyzes phosphate transfer from ATP to guanidoacetate, and arginine kinase catalyzes phosphate transfer from ATP to argenine. A cysteine-containing active site is conserved in this family (PROSITE: PDOC00103).

Prenyl transferases are heterodimers, consisting of an alpha and a beta subunit, that catalyze the transfer of an isoprenyl group. An example of a prenyl transferase is the mammalian protein farnesyl transferase. The alpha subunit of farnesyl transferase consists of 5 repeats of 34 amino acids each, with each repeat containing an invariant tryptophan (PROSITE: PDOC00703).

Saccharyl transferases are glycating enzymes involved in a variety of metabolic processes. Oligosacchryl transferase-48, for example, is a receptor for advanced glycation endproducts. Accumulation of these endproducts is observed in vascular complications of diabetes, macrovascular disease, renal insufficiency, and Alzheimer's disease (Thornalley, P.J. (1998) Cell Mol. Biol. (Noisy-Le-Grand) 44:1013-1023).

Coenzyme A (CoA) transferase catalyzes the transfer of CoA between two carboxylic acids. Succinyl CoA:3-oxoacid CoA transferase, for example, transfers CoA from succinyl-CoA to a recipient such as acetoacetate. Acetoacetate is essential to the metabolism of ketone bodies, which accumulate in tissues affected by metabolic disorders such as diabetes (PROSITE: PDOC00980).

Hydrolases

Hydrolysis is the breaking of a covalent bond in a substrate by introduction of a molecule of water. The reaction involves a nucleophilic attack by the water molecule's oxygen atom on a target bond in the substrate. The water molecule is split across the target bond, breaking the bond and generating two product molecules. Hydrolases participate in reactions essential to such functions as synthesis and degradation of cell components, and for regulation of cell functions including cell

signaling, cell proliferation, inflamation, apoptosis, secretion and excretion. Hydrolases are involved in key steps in disease processes involving these functions. Hydrolytic enzymes, or hydrolases, may be grouped by substrate specificity into classes including phosphatases, peptidases, lysophospholipases, phosphodiesterases, glycosidases, and glyoxalases.

Phosphatases hydrolytically remove phosphate groups from proteins, an energy-providing step that regulates many cellular processes, including intracellular signaling pathways that in turn control cell growth and differentiation, cell-cell contact, the cell cycle, and oncogenesis.

Lysophospholipases (LPLs) regulate intracellular lipids by catalyzing the hydrolysis of ester bonds to remove an acyl group, a key step in lipid degradation. Small LPL isoforms, approximately 15-30 kD, function as hydrolases; larger isoforms function both as hydrolases and transacylases. A particular substrate for LPLs, lysophosphatidylcholine, causes lysis of cell membranes. LPL activity is regulated by signaling molecules important in numerous pathways, including the inflammatory response.

Peptidases, also called proteases, cleave peptide bonds that form the backbone of peptide or protein chains. Proteolytic processing is essential to cell growth, differentiation, remodeling, and homeostasis as well as inflammation and immune response. Since typical protein half-lives range from hours to a few days, peptidases are continually cleaving precursor proteins to their active form, removing signal sequences from targeted proteins, and degrading aged or defective proteins. Peptidases function in bacterial, parasitic, and viral invasion and replication within a host. Examples of peptidases include trypsin and chymotrypsin (components of the complement cascade and the blood-clotting cascade) lysosomal cathepsins, calpains, pepsin, renin, and chymosin (Beynon, R.J. and J.S. Bond (1994) Proteolytic Enzymes: A Practical Approach, Oxford University Press, New York NY, pp. 1-5).

The phosphodiesterases catalyze the hydrolysis of one of the two ester bonds in a phosphodiester compound. Phosphodiesterases are therefore crucial to a variety of cellular processes. Phosphodiesterases include DNA and RNA endo- and exo-nucleases, which are essential to cell growth and replication as well as protein synthesis. Another phosphodiesterase is acid sphingomyelinase, which hydrolyzes the membrane phospholipid sphingomyelin to ceramide and phosphorylcholine. Phosphorylcholine is used in the synthesis of phosphatidylcholine, which is involved in numerous intracellular signaling pathways. Ceramide is an essential precursor for the generation of gangliosides, membrane lipids found in high concentration in neural tissue. Defective acid sphingomyelinase phosphodiesterase leads to a build-up of sphingomyelin molecules in lysosomes, resulting in Niemann-Pick disease.

Glycosidases catalyze the cleavage of hemiacetyl bonds of glycosides, which are compounds that contain one or more sugar. Mammalian lactase-phlorizin hydrolase, for example, is an intestinal

enzyme that splits lactose. Mammalian beta-galactosidase removes the terminal galactose from gangliosides, glycoproteins, and glycosaminoglycans, and deficiency of this enzyme is associated with a gangliosidosis known as Morquio disease type B. Vertebrate lysosomal alpha-glucosidase, which hydrolyzes glycogen, maltose, and isomaltose, and vertebrate intestinal sucrase-isomaltase, which hydrolyzes sucrose, maltose, and isomaltose, are widely distributed members of this family with highly conserved sequences at their active sites.

The glyoxylase system is involved in gluconeogenesis, the production of glucose from storage compounds in the body. It consists of glyoxylase I, which catalyzes the formation of S-D-lactoylglutathione from methyglyoxal, a side product of triose-phosphate energy metabolism, and glyoxylase II, which hydrolyzes S-D-lactoylglutathione to D-lactic acid and reduced glutathione. Glyoxylases are involved in hyperglycemia, non-insulin-dependent diabetes mellitus, the detoxification of bacterial toxins, and in the control of cell proliferation and microtubule assembly.

Lyases

Lyases are a class of enzymes that catalyze the cleavage of C-C, C-O, C-N, C-S, C-(halide), P-O or other bonds without hydrolysis or oxidation to form two molecules, at least one of which contains a double bond (Stryer, L. (1995) Biochemistry W.H. Freeman and Co. New York, NY p.620). Lyases are critical components of cellular biochemistry with roles in metabolic energy production including fatty acid metabolism, as well as other diverse enzymatic processes. Further classification of lyases reflects the type of bond cleaved as well as the nature of the cleaved group.

The group of C-C lyases include carboxyl-lyases (decarboxylases), aldehyde-lyases (aldolases), oxo-acid-lyases and others. The C-O lyase group includes hydro-lyases, lyases acting on polysaccharides and other lyases. The C-N lyase group includes ammonia-lyases, amidine-lyases, amine-lyases (deaminases) and other lyases.

Proper regulation of lyases is critical to normal physiology. For example, mutation induced deficiencies in the uroporphyrinogen decarboxylase can lead to photosensitive cutaneous lesions in the genetically-linked disorder familial porphyria cutanea tarda (Mendez, M. et al. (1998) Am. J. Genet. 63:1363-1375). It has also been shown that adenosine deaminase (ADA) deficiency stems from genetic mutations in the ADA gene, resulting in the disorder severe combined immunodeficiency disease (SCID) (Hershfield, M.S. (1998) Semin. Hematol. 35:291-298).

Isomerases

Isomerases are a class of enzymes that catalyze geometric or structural changes within a molecule to form a single product. This class includes racemases and epimerases, cis-trans-isomerases, intramolecular oxidoreductases, intramolecular transferases (mutases) and intramolecular lyases. Isomerases are critical components of cellular biochemistry with roles in metabolic energy production including glycolysis, as well as other diverse enzymatic processes (Stryer, L. (1995)

Biochemistry, W.H. Freeman and Co., New York NY, pp.483-507).

Racemases are a subset of isomerases that catalyze inversion of a molecules configuration around the asymmetric carbon atom in a substrate having a single center of asymmetry, thereby interconverting two racemers. Epimerases are another subset of isomerases that catalyze inversion of configuration around an asymmetric carbon atom in a substrate with more than one center of symmetry, thereby interconverting two epimers. Racemases and epimerases can act on amino acids and derivatives, hydroxy acids and derivatives, as well as carbohydrates and derivatives. The interconversion of UDP-galactose and UDP-glucose is catalyzed by UDP-galactose-4'-epimerase. Proper regulation and function of this epimerase is essential to the synthesis of glycoproteins and glycolipids. Elevated blood galactose levels have been correlated with UDP-galactose-4'-epimerase deficiency in screening programs of infants (Gitzelmann, R. (1972) Helv. Paediat. Acta 27:125-130).

Oxidoreductases can be isomerases as well. Oxidoreductases catalyze the reversible transfer of electrons from a substrate that becomes oxidized to a substrate that becomes reduced. This class of enzymes includes dehydrogenases, hydroxylases, oxidases, oxygenases, peroxidases, and reductases. Proper maintenance of oxidoreductase levels is physiologically important. For example, genetically-linked deficiencies in lipoamide dehydrogenase can result in lactic acidosis (Robinson, B.H. et al. (1977) Pediat. Res. 11:1198-1202).

Another subgroup of isomerases are the transferases (or mutases). Transferases transfer a chemical group from one compound (the donor) to another compound (the acceptor). The types of groups transferred by these enzymes include acyl groups, amino groups, phosphate groups (phosphotransferases or phosphomutases), and others. The transferase carnitine palmitoyltransferase is an important component of fatty acid metabolism. Genetically-linked deficiencies in this transferase can lead to myopathy (Scriver, C.R. et al. (1995) The Metabolic and Molecular Basis of Inherited Disease, McGraw-Hill, New York NY, pp.1501-1533).

Yet another subgroup of isomerases are the topoisomersases. Topoisomerases are enzymes that affect the topological state of DNA. For example, defects in topoisomerases or their regulation can affect normal physiology. Reduced levels of topoisomerase II have been correlated with some of the DNA processing defects associated with the disorder ataxia-telangiectasia (Singh, S.P. et al. (1988) Nucleic Acids Res. 16:3919-3929).

Ligases

Ligases catalyze the formation of a bond between two substrate molecules. The process involves the hydrolysis of a pyrophosphate bond in ATP or a similar energy donor. Ligases are classified based on the nature of the type of bond they form, which can include carbon-oxygen, carbon-sulfur, carbon-nitrogen, carbon-carbon and phosphoric ester bonds.

Ligases forming carbon-oxygen bonds include the aminoacyl-transfer RNA (tRNA)

synthetases which are important RNA-associated enzymes with roles in translation. Protein biosynthesis depends on each amino acid forming a linkage with the appropriate tRNA. The aminoacyl-tRNA synthetases are responsible for the activation and correct attachment of an amino acid with its cognate tRNA. The 20 aminoacyl-tRNA synthetase enzymes can be divided into two structural classes, and each class is characterized by a distinctive topology of the catalytic domain. Class I enzymes contain a catalytic domain based on the nucleotide-binding Rossman fold. Class II enzymes contain a central catalytic domain, which consists of a seven-stranded antiparallel ß-sheet motif, as well as N- and C- terminal regulatory domains. Class II enzymes are separated into two groups based on the heterodimeric or homodimeric structure of the enzyme; the latter group is further subdivided by the structure of the N- and C-terminal regulatory domains (Hartlein, M. and S. Cusack (1995) J. Mol. Evol. 40:519-530). Autoantibodies against aminoacyl-tRNAs are generated by patients with dermatomyositis and polymyositis, and correlate strongly with complicating interstitial lung disease (ILD). These antibodies appear to be generated in response to viral infection, and coxsackie virus has been used to induce experimental viral myositis in animals.

Ligases forming carbon-sulfur bonds (Acid-thiol ligases) mediate a large number of cellular biosynthetic intermediary metabolism processes involve intermolecular transfer of carbon atom-containing substrates (carbon substrates). Examples of such reactions include the tricarboxylic acid cycle, synthesis of fatty acids and long-chain phospholipids, synthesis of alcohols and aldehydes, synthesis of intermediary metabolites, and reactions involved in the amino acid degradation pathways. Some of these reactions require input of energy, usually in the form of conversion of ATP to either ADP or AMP and pyrophosphate.

In many cases, a carbon substrate is derived from a small molecule containing at least two carbon atoms. The carbon substrate is often covalently bound to a larger molecule which acts as a carbon substrate carrier molecule within the cell. In the biosynthetic mechanisms described above, the carrier molecule is coenzyme A. Coenzyme A (CoA) is structurally related to derivatives of the nucleotide ADP and consists of 4'-phosphopantetheine linked via a phosphodiester bond to the alpha phosphate group of adenosine 3',5'-bisphosphate. The terminal thiol group of 4'-phosphopantetheine acts as the site for carbon substrate bond formation. The predominant carbon substrates which utilize CoA as a carrier molecule during biosynthesis and intermediary metabolism in the cell are acetyl, succinyl, and propionyl moieties, collectively referred to as acyl groups. Other carbon substrates include enoyl lipid, which acts as a fatty acid oxidation intermediate, and carnitine, which acts as an acetyl-CoA flux regulator/ mitochondrial acyl group transfer protein. Acyl-CoA and acetyl-CoA are synthesized in the cell by acyl-CoA synthetase and acetyl-CoA synthetase, respectively.

Activation of fatty acids is mediated by at least three forms of acyl-CoA synthetase activity: i) acetyl-CoA synthetase, which activates acetate and several other low molecular weight carboxylic

acids and is found in muscle mitochondria and the cytosol of other tissues; ii) medium-chain acyl-CoA synthetase, which activates fatty acids containing between four and eleven carbon atoms (predominantly from dietary sources), and is present only in liver mitochondria; and iii) acyl CoA synthetase, which is specific for long chain fatty acids with between six and twenty carbon atoms, and is found in microsomes and the mitochondria. Proteins associated with acyl-CoA synthetase activity have been identified from many sources including bacteria, yeast, plants, mouse, and man. The activity of acyl-CoA synthetase may be modulated by phosphorylation of the enzyme by cAMP-dependent protein kinase.

Ligases forming carbon-nitrogen bonds include amide synthases such as glutamine synthetase (glutamate-ammonia ligase) that catalyzes the amination of glutamic acid to glutamine by ammonia using the energy of ATP hydrolysis. Glutamine is the primary source for the amino group in various amide transfer reactions involved in de novo pyrimidine nucleotide synthesis and in purine and pyrimidine ribonucleotide interconversions. Overexpression of glutamine synthetase has been observed in primary liver cancer (Christa, L. et al. (1994) Gastroent. 106:1312-1320).

Acid-amino-acid ligases (peptide synthases) are represented by the ubiquitin proteases which are associated with the ubiquitin conjugation system (UCS), a major pathway for the degradation of cellular proteins in eukaryotic cells and some bacteria. The UCS mediates the elimination of abnormal proteins and regulates the half-lives of important regulatory proteins that control cellular processes such as gene transcription and cell cycle progression. In the UCS pathway, proteins targeted for degradation are conjugated to a ubiquitin (Ub), a small heat stable protein. Ub is first activated by a ubiquitin-activating enzyme (E1), and then transferred to one of several Ub-conjugating enzymes (E2). E2 then links the Ub molecule through its C-terminal glycine to an internal lysine (acceptor lysine) of a target protein. The ubiquitinated protein is then recognized and degraded by proteasome, a large, multisubunit proteolytic enzyme complex, and ubiquitin is released for reutilization by ubiquitin protease. The UCS is implicated in the degradation of mitotic cyclic kinases, oncoproteins, tumor suppressor genes such as p53, viral proteins, cell surface receptors associated with signal transduction, transcriptional regulators, and mutated or damaged proteins (Ciechanover, A. (1994) Cell 79:13-21). A murine proto-oncogene, Unp, encodes a nuclear ubiquitin protease whose overexpression leads to oncogenic transformation of NIH3T3 cells, and the human homolog of this gene is consistently elevated in small cell tumors and adenocarcinomas of the lung (Gray, D.A. (1995) Oncogene 10:2179-2183).

Cyclo-ligases and other carbon-nitrogen ligases comprise various enzymes and enzyme complexes that participate in the de novo pathways to purine and pyrimidine biosynthesis. Because these pathways are critical to the synthesis of nucleotides for replication of both RNA and DNA, many of these enzymes have been the targets of clinical agents for the treatment of cell proliferative

disorders such as cancer and infectious diseases.

Purine biosynthesis occurs de novo from the amino acids glycine and glutamine, and other small molecules. Three of the key reactions in this process are catalyzed by a trifunctional enzyme composed of glycinamide-ribonucleotide synthetase (GARS), aminoimidazole ribonucleotide synthetase (AIRS), and glycinamide ribonucleotide transformylase (GART). Together these three enzymes combine ribosylamine phosphate with glycine to yield phosphoribosyl aminoimidazole, a precursor to both adenylate and guanylate nucleotides. This trifunctional protein has been implicated in the pathology of Downs syndrome (Aimi, J. et al. (1990) Nucleic Acid Res. 18:6665-6672). Adenylosuccinate synthetase catalyzes a later step in purine biosynthesis that converts inosinic acid to adenylosuccinate, a key step on the path to ATP synthesis. This enzyme is also similar to another carbon-nitrogen ligase, argininosuccinate synthetase, that catalyzes a similar reaction in the urea cycle (Powell, S.M. et al. (1992) FEBS Lett. 303:4-10).

Like the de novo biosynthesis of purines, de novo synthesis of the pyrimidine nucleotides uridylate and cytidylate also arises from a common precursor, in this instance the nucleotide orotidylate derived from orotate and phosphoribosyl pyrophosphate (PPRP). Again a trifunctional enzyme comprising three carbon-nitrogen ligases plays a key role in the process. In this case the enzymes aspartate transcarbamylase (ATCase), carbamyl phosphate synthetase II, and dihydroorotase (DHOase) are encoded by a single gene called CAD. Together these three enzymes combine the initial reactants in pyrimidine biosynthesis, glutamine, $CO_2$, and ATP to form dihydroorotate, the precursor to orotate and orotidylate (Iwahana, H. et al. (1996) Biochem. Biophys. Res. Commun. 219:249-255). Further steps then lead to the synthesis of uridine nucleotides from orotidylate. Cytidine nucleotides are derived from uridine-5'-triphosphate (UTP) by the amidation of UTP using glutamine as the amino donor and the enzyme CTP synthetase. Regulatory mutations in the human CTP synthetase are believed to confer multi-drug resistance to agents widely used in cancer therapy (Yamauchi, M. et al. (1990) EMBO J. 9:2095-2099).

Ligases forming carbon-carbon bonds include the carboxylases acetyl-CoA carboxylase and pyruvate carboxylase. Acetyl-CoA carboxylase catalyzes the carboxylation of acetyl-CoA from $CO_2$ and $H_2O$ using the energy of ATP hydrolysis. Acetyl-CoA carboxylase is the rate-limiting step in the biogenesis of long-chain fatty acids. Two isoforms of acetyl-CoA carboxylase, types I and types II, are expressed in human in a tissue-specific manner (Ha, J. et al. (1994) Eur. J. Biochem. 219:297-306). Pyruvate carboxylase is a nuclear-encoded mitochondrial enzyme that catalyzes the conversion of pyruvate to oxaloacetate, a key intermediate in the citric acid cycle.

Ligases forming phosphoric ester bonds include the DNA ligases involved in both DNA replication and repair. DNA ligases seal phosphodiester bonds between two adjacent nucleotides in a DNA chain using the energy from ATP hydrolysis to first activate the free 5'-phosphate of one

nucleotide and then react it with the 3'-OH group of the adjacent nucleotide. This resealing reaction is used in both DNA replication to join small DNA fragments called Okazaki fragments that are transiently formed in the process of replicating new DNA, and in DNA repair. DNA repair is the process by which accidental base changes, such as those produced by oxidative damage, hydrolytic attack, or uncontrolled methylation of DNA, are corrected before replication or transcription of the DNA can occur. Bloom's syndrome is an inherited human disease in which individuals are partially deficient in DNA ligation and consequently have an increased incidence of cancer (Alberts, B. et al. (1994) The Molecular Biology of the Cell, Garland Publishing Inc., New York NY, p. 247).

**Molecules Associated with Growth and Development**

SEQ ID NO:51 and SEQ ID NO:52 encode, for example, molecules associated with growth and development.

Human growth and development requires the spatial and temporal regulation of cell differentiation, cell proliferation, and apoptosis. These processes coordinately control reproduction, aging, embryogenesis, morphogenesis, organogenesis, and tissue repair and maintenance. At the cellular level, growth and development is governed by the cell's decision to enter into or exit from the cell division cycle and by the cell's commitment to a terminally differentiated state. These decisions are made by the cell in response to extracellular signals and other environmental cues it receives. The following discussion focuses on the molecular mechanisms of cell division, reproduction, cell differentiation and proliferation, apoptosis, and aging.

Cell Division

Cell division is the fundamental process by which all living things grow and reproduce. In unicellular organisms such as yeast and bacteria, each cell division doubles the number of organisms, while in multicellular species many rounds of cell division are required to replace cells lost by wear or by programmed cell death, and for cell differentiation to produce a new tissue or organ. Details of the cell division cycle may vary, but the basic process consists of three principle events. The first event, interphase, involves preparations for cell division, replication of the DNA, and production of essential proteins. In the second event, mitosis, the nuclear material is divided and separates to opposite sides of the cell. The final event, cytokinesis, is division and fission of the cell cytoplasm. The sequence and timing of cell cycle transitions is under the control of the cell cycle regulation system which controls the process by positive or negative regulatory circuits at various check points.

Regulated progression of the cell cycle depends on the integration of growth control pathways with the basic cell cycle machinery. Cell cycle regulators have been identified by selecting for human and yeast cDNAs that block or activate cell cycle arrest signals in the yeast mating pheromone pathway when they are overexpressed. Known regulators include human CPR (cell cycle progression

restoration) genes, such as CPR8 and CPR2, and yeast CDC (cell division control) genes, including CDC91, that block the arrest signals. The CPR genes express a variety of proteins including cyclins, tumor suppressor binding proteins, chaperones, transcription factors, translation factors, and RNA-binding proteins (Edwards, M.C. et al.(1997) Genetics 147:1063-1076).

Several cell cycle transitions, including the entry and exit of a cell from mitosis, are dependent upon the activation and inhibition of cyclin-dependent kinases (Cdks). The Cdks are composed of a kinase subunit, Cdk, and an activating subunit, cyclin, in a complex that is subject to many levels of regulation. There appears to be a single Cdk in Saccharomyces cerevisiae and Saccharomyces pombe whereas mammals have a variety of specialized Cdks. Cyclins act by binding to and activating cyclin-dependent protein kinases which then phosphorylate and activate selected proteins involved in the mitotic process. The Cdk-cyclin complex is both positively and negatively regulated by phosphorylation, and by targeted degradation involving molecules such as CDC4 and CDC53. In addition, Cdks are further regulated by binding to inhibitors and other proteins such as Suc1 that modify their specificity or accessibility to regulators (Patra, D. and W.G. Dunphy (1996) Genes Dev. 10:1503-1515; and Mathias, N. et al. (1996) Mol. Cell Biol. 16:6634-6643).

Reproduction

The male and female reproductive systems are complex and involve many aspects of growth and development. The anatomy and physiology of the male and female reproductive systems are reviewed in (Guyton, A.C. (1991) Textbook of Medical Physiology, W.B. Saunders Co., Philadelphia PA, pp. 899-928).

The male reproductive system includes the process of spermatogenesis, in which the sperm are formed, and male reproductive functions are regulated by various hormones and their effects on accessory sexual organs, cellular metabolism, growth, and other bodily functions.

Spermatogenesis begins at puberty as a result of stimulation by gonadotropic hormones released from the anterior pituitary. Immature sperm (spermatogonia) undergo several mitotic cell divisions before undergoing meiosis and full maturation. The testes secrete several male sex hormones, the most abundant being testosterone, that is essential for growth and division of the immature sperm, and for the masculine characteristics of the male body. Three other male sex hormones, gonadotropin-releasing hormone (GnRH), luteinizing hormone (LH), and follicle-stimulating hormone (FSH) control sexual function.

The uterus, ovaries, fallopian tubes, vagina, and breasts comprise the female reproductive system. The ovaries and uterus are the source of ova and the location of fetal development, respectively. The fallopian tubes and vagina are accessory organs attached to the top and bottom of the uterus, respectively. Both the uterus and ovaries have additional roles in the development and loss of reproductive capability during a female's lifetime. The primary role of the breasts is lactation.

Multiple endocrine signals from the ovaries, uterus, pituitary, hypothalamus, adrenal glands, and other tissues coordinate reproduction and lactation. These signals vary during the monthly menstruation cycle and during the female's lifetime. Similarly, the sensitivity of reproductive organs to these endocrine signals varies during the female's lifetime.

A combination of positive and negative feedback to the ovaries, pituitary and hypothalamus glands controls physiologic changes during the monthly ovulation and endometrial cycles. The anterior pituitary secretes two major gonadotropin hormones, follicle-stimulating hormone (FSH) and luteinizing hormone (LH), regulated by negative feedback of steroids, most notably by ovarian estradiol. If fertilization does not occur, estrogen and progesterone levels decrease. This sudden reduction of the ovarian hormones leads to menstruation, the desquamation of the endometrium.

Hormones further govern all the steps of pregnancy, parturition, lactation, and menopause. During pregnancy large quantities of human chorionic gonadotropin (hCG), estrogens, progesterone, and human chorionic somatomammotropin (hCS) are formed by the placenta. hCG, a glycoprotein similar to luteinizing hormone, stimulates the corpus luteum to continue producing more progesterone and estrogens, rather than to involute as occurs if the ovum is not fertilized. hCS is similar to growth hormone and is crucial for fetal nutrition.

The female breast also matures during pregnancy. Large amounts of estrogen secreted by the placenta trigger growth and branching of the breast milk ductal system while lactation is initiated by the secretion of prolactin by the pituitary gland.

Parturition involves several hormonal changes that increase uterine contractility toward the end of pregnancy, as follows. The levels of estrogens increase more than those of progesterone. Oxytocin is secreted by the neurohypophysis. Concomitantly, uterine sensitivity to oxytocin increases. The fetus itself secretes oxytocin, cortisol (from adrenal glands), and prostaglandins.

Menopause occurs when most of the ovarian follicles have degenerated. The ovary then produces less estradiol, reducing the negative feedback on the pituitary and hypothalamus glands. Mean levels of circulating FSH and LH increase, even as ovulatory cycles continue. Therefore, the ovary is less responsive to gonadotropins, and there is an increase in the time between menstrual cycles. Consequently, menstrual bleeding ceases and reproductive capability ends.

Cell Differentiation and Proliferation

Tissue growth involves complex and ordered patterns of cell proliferation, cell differentiation, and apoptosis. Cell proliferation must be regulated to maintain both the number of cells and their spatial organization. This regulation depends upon the appropriate expression of proteins which control cell cycle progression in response to extracellular signals, such as growth factors and other mitogens, and intracellular cues, such as DNA damage or nutrient starvation. Molecules which directly or indirectly modulate cell cycle progression fall into several categories, including growth factors and their

receptors, second messenger and signal transduction proteins, oncogene products, tumor-suppressor proteins, and mitosis-promoting factors.

Growth factors were originally described as serum factors required to promote cell proliferation. Most growth factors are large, secreted polypeptides that act on cells in their local environment. Growth factors bind to and activate specific cell surface receptors and initiate intracellular signal transduction cascades. Many growth factor receptors are classified as receptor tyrosine kinases which undergo autophosphorylation upon ligand binding. Autophosphorylation enables the receptor to interact with signal transduction proteins characterized by the presence of SH2 or SH3 domains (Src homology regions 2 or 3). These proteins then modulate the activity state of small G-proteins, such as Ras, Rab, and Rho, along with GTPase activating proteins (GAPs), guanine nucleotide releasing proteins (GNRPs), and other guanine nucleotide exchange factors. Small G proteins act as molecular switches that activate other downstream events, such as mitogen-activated protein kinase (MAP kinase) cascades. MAP kinases ultimately activate transcription of mitosis-promoting genes.

In addition to growth factors, small signaling peptides and hormones also influence cell proliferation. These molecules bind primarily to another class of receptor, the trimeric G-protein coupled receptor (GPCR), found predominantly on the surface of immune, neuronal and neuroendocrine cells. Upon ligand binding, the GPCR activates a trimeric G protein which in turn triggers increased levels of intracellular second messengers such as phospholipase C, Ca2+, and cyclic AMP. Most GPCR-mediated signaling pathways indirectly promote cell proliferation by causing the secretion or breakdown of other signaling molecules that have direct mitogenic effects. These signaling cascades often involve activation of kinases and phosphatases. Some growth factors, such as some members of the transforming growth factor beta (TGF-β) family, act on some cells to stimulate cell proliferation and on other cells to inhibit it. Growth factors may also stimulate a cell at one concentration and inhibit the same cell at another concentration. Most growth factors also have a multitude of other actions besides the regulation of cell growth and division: they can control the proliferation, survival, differentiation, migration, or function of cells depending on the circumstance. For example, the tumor necrosis factor/nerve growth factor (TNF/NGF) family can activate or inhibit cell death, as well as regulate proliferation and differentiation. The cell response depends on the type of cell, its stage of differentiation and transformation status, which surface receptors are stimulated, and the types of stimuli acting on the cell (Smith, A. et al. (1994) Cell 76:959-962; and Nocentini, G. et al. (1997) Proc. Natl. Acad. Sci. USA 94:6216-6221).

Neighboring cells in a tissue compete for growth factors, and when provided with "unlimited" quantities in a perfused system will grow to even higher cell densities before reaching density-dependent inhibition of cell division. Cells often demonstrate an anchorage dependence of cell division as well.

This anchorage dependence may be associated with the formation of focal contacts linking the cytoskeleton with the extracellular matrix (ECM). The expression of ECM components can be stimulated by growth factors. For example, TGF-β stimulates fibroblasts to produce a variety of ECM proteins, including fibronectin, collagen, and tenascin (Pearson, C.A. et al. (1988) EMBO J. 7:2677-2981). In fact, for some cell types specific ECM molecules, such as laminin or fibronectin, may act as growth factors. Tenascin-C and -R, expressed in developing and lesioned neural tissue, provide stimulatory/anti-adhesive or inhibitory properties, respectively, for axonal growth (Faissner, A. (1997) Cell Tissue Res. 290:331-341).

Cancers are associated with the activation of oncogenes which are derived from normal cellular genes. These oncogenes encode oncoproteins which convert normal cells into malignant cells. Some oncoproteins are mutant isoforms of the normal protein, and other oncoproteins are abnormally expressed with respect to location or amount of expression. The latter category of oncoprotein causes cancer by altering transcriptional control of cell proliferation. Five classes of oncoproteins are known to affect cell cycle controls. These classes include growth factors, growth factor receptors, intracellular signal transducers, nuclear transcription factors, and cell-cycle control proteins. Viral oncogenes are integrated into the human genome after infection of human cells by certain viruses. Examples of viral oncogenes include v-src, v-abl, and v-fps.

Many oncogenes have been identified and characterized. These include sis, erbA, erbB, her-2, mutated $G_s$, src, abl, ras, crk, jun, fos, myc, and mutated tumor-suppressor genes such as RB, p53, mdm2, Cip1, p16, and cyclin D. Transformation of normal genes to oncogenes may also occur by chromosomal translocation. The Philadelphia chromosome, characteristic of chronic myeloid leukemia and a subset of acute lymphoblastic leukemias, results from a reciprocal translocation between chromosomes 9 and 22 that moves a truncated portion of the proto-oncogene c-abl to the breakpoint cluster region (bcr) on chromosome 22.

Tumor-suppressor genes are involved in regulating cell proliferation. Mutations which cause reduced or loss of function in tumor-suppressor genes result in uncontrolled cell proliferation. For example, the retinoblastoma gene product (RB), in a non-phosphorylated state, binds several early-response genes and suppresses their transcription, thus blocking cell division. Phosphorylation of RB causes it to dissociate from the genes, releasing the suppression, and allowing cell division to proceed.

Apoptosis

Apoptosis is the genetically controlled process by which unneeded or defective cells undergo programmed cell death. Selective elimination of cells is as important for morphogenesis and tissue remodeling as is cell proliferation and differentiation. Lack of apoptosis may result in hyperplasia and other disorders associated with increased cell proliferation. Apoptosis is also a critical component of the immune response. Immune cells such as cytotoxic T-cells and natural killer cells prevent the spread

of disease by inducing apoptosis in tumor cells and virus-infected cells. In addition, immune cells that fail to distinguish self molecules from foreign molecules must be eliminated by apoptosis to avoid an autoimmune response.

Apoptotic cells undergo distinct morphological changes. Hallmarks of apoptosis include cell shrinkage, nuclear and cytoplasmic condensation, and alterations in plasma membrane topology. Biochemically, apoptotic cells are characterized by increased intracellular calcium concentration, fragmentation of chromosomal DNA, and expression of novel cell surface components.

The molecular mechanisms of apoptosis are highly conserved, and many of the key protein regulators and effectors of apoptosis have been identified. Apoptosis generally proceeds in response to a signal which is transduced intracellularly and results in altered patterns of gene expression and protein activity. Signaling molecules such as hormones and cytokines are known both to stimulate and to inhibit apoptosis through interactions with cell surface receptors. Transcription factors also play an important role in the onset of apoptosis. A number of downstream effector molecules, particularly proteases such as the cysteine proteases called caspases, have been implicated in the degradation of cellular components and the proteolytic activation of other apoptotic effectors.

Aging and Senescence

Studies of the aging process or senescence have shown a number of characteristic cellular and molecular changes (Fauci et al. (1998) Harrison's Principles of Internal Medicine, McGraw-Hill, New York NY, p.37). These characteristics include increases in chromosome structural abnormalities, DNA cross-linking, incidence of single-stranded breaks in DNA, losses in DNA methylation, and degradation of telomere regions. In addition to these DNA changes, post-translational alterations of proteins increase including, deamidation, oxidation, cross-linking, and nonenzymatic glycation. Still further molecular changes occur in the mitochondria of aging cells through deterioration of structure. These changes eventually contribute to decreased function in every organ of the body.

**Biochemical Pathway Molecules**

SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, and SEQ ID NO:50 encode, for example, biochemical pathway molecules.

Biochemical pathways are responsible for regulating metabolism, growth and development, protein secretion and trafficking, environmental responses, and ecological interactions including immune response and response to parasites.

DNA replication

Deoxyribonucleic acid (DNA), the genetic material, is found in both the nucleus and mitochondria of human cells. The bulk of human DNA is nuclear, in the form of linear chromosomes, while mitochondrial DNA is circular. DNA replication begins at specific sites called origins of

replication. Bidirectional synthesis occurs from the origin via two growing forks that move in opposite directions. Replication is semi-conservative, with each daughter duplex containing one old strand and its newly synthesized complementary partner. Proteins involved in DNA replication include DNA polymerases, DNA primase, telomerase, DNA helicase, topoisomerases, DNA ligases, replication factors, and DNA-binding proteins.

DNA Recombination and Repair

Cells are constantly faced with replication errors and environmental assault (such as ultraviolet irradiation) that can produce DNA damage. Damage to DNA consists of any change that modifies the structure of the molecule. Changes to DNA can be divided into two general classes, single base changes and structural distortions. Any damage to DNA can produce a mutation, and the mutation may produce a disorder, such as cancer.

Changes in DNA are recognized by repair systems within the cell. These repair systems act to correct the damage and thus prevent any deleterious affects of a mutational event. Repair systems can be divided into three general types, direct repair, excision repair, and retrieval systems. Proteins involved in DNA repair include DNA polymerase, excision repair proteins, excision and cross link repair proteins, recombination and repair proteins, RAD51 proteins, and BLN and WRN proteins that are homologs of RecQ helicase. When the repair systems are eliminated, cells become exceedingly sensitive to environmental mutagens, such as ultraviolet irradiation. Patients with disorders associated with a loss in DNA repair systems often exhibit a high sensitivity to environmental mutagens. Examples of such disorders include xeroderma pigmentosum (XP), Bloom's syndrome (BS), and Werner's syndrome (WS) (Yamagata, K. et al. (1998) Proc. Natl. Acad. Sci. USA 95:8733-8738), ataxia telangiectasia, Cockayne's syndrome, and Fanconi's anemia.

Recombination is the process whereby new DNA sequences are generated by the movements of large pieces of DNA. In homologous recombination, which occurs during meiosis and DNA repair, parent DNA duplexes align at regions of sequence similarity, and new DNA molecules form by the breakage and joining of homologous segments. Proteins involved include RAD51 recombinase. In site-specific recombination, two specific but not necessarily homologous DNA sequences are exchanged. In the immune system this process generates a diverse collection of antibody and T cell receptor genes. Proteins involved in site-specific recombination in the immune system include recombination activating genes 1 and 2 (RAG1 and RAG2). A defect in immune system site-specific recombination causes severe combined immunodeficiency disease in mice.

RNA Metabolism

Ribonucleic acid (RNA) is a linear single-stranded polymer of four nucleotides, ATP, CTP, UTP, and GTP. In most organisms, RNA is transcribed as a copy of DNA, the genetic material of the organism. In retroviruses RNA rather than DNA serves as the genetic material. RNA copies of the

genetic material encode proteins or serve various structural, catalytic, or regulatory roles in organisms. RNA is classified according to its cellular localization and function. Messenger RNAs (mRNAs) encode polypeptides. Ribosomal RNAs (rRNAs) are assembled, along with ribosomal proteins, into ribosomes, which are cytoplasmic particles that translate mRNA into polypeptides. Transfer RNAs (tRNAs) are cytosolic adaptor molecules that function in mRNA translation by recognizing both an mRNA codon and the amino acid that matches that codon. Heterogeneous nuclear RNAs (hnRNAs) include mRNA precursors and other nuclear RNAs of various sizes. Small nuclear RNAs (snRNAs) are a part of the nuclear spliceosome complex that removes intervening, non-coding sequences (introns) and rejoins exons in pre-mRNAs.

RNA Transcription

The transcription process synthesizes an RNA copy of DNA. Proteins involved include multi-subunit RNA polymerases, transcription factors IIA, IIB, IID, IIE, IIF, IIH, and IIJ. Many transcription factors incorporate DNA-binding structural motifs which comprise either α-helices or β-sheets that bind to the major groove of DNA. Four well-characterized structural motifs are helix-turn-helix, zinc finger, leucine zipper, and helix-loop-helix.

RNA Processing

Various proteins are necessary for processing of transcribed RNAs in the nucleus. Pre-mRNA processing steps include capping at the 5' end with methylguanosine, polyadenylating the 3' end, and splicing to remove introns. The spliceosomal complex is comprised of five small nuclear ribonucleoprotein particles (snRNPs) designated U1, U2, U4, U5, and U6. Each snRNP contains a single species of snRNA and about ten proteins. The RNA components of some snRNPs recognize and base-pair with intron consensus sequences. The protein components mediate spliceosome assembly and the splicing reaction. Autoantibodies to snRNP proteins are found in the blood of patients with systemic lupus erythematosus (Stryer, L. (1995) Biochemistry W.H. Freeman and Company, New York NY, p. 863).

Heterogeneous nuclear ribonucleoproteins (hnRNPs) have been identified that have roles in splicing, exporting of the mature RNAs to the cytoplasm, and mRNA translation (Biamonti, G. et al. (1998) Clin. Exp. Rheumatol. 16:317-326). Some examples of hnRNPs include the yeast proteins Hrp1p, involved in cleavage and polyadenylation at the 3' end of the RNA; Cbp80p, involved in capping the 5' end of the RNA; and Npl3p, a homolog of mammalian hnRNP A1, involved in export of mRNA from the nucleus (Shen, E.C. et al. (1998) Genes Dev. 12:679-691). HnRNPs have been shown to be important targets of the autoimmune response in rheumatic diseases (Biamonti, supra).

Many snRNP proteins, hnRNP proteins, and alternative splicing factors are characterized by an RNA recognition motif (RRM). (Reviewed in Birney, E. et al. (1993) Nucleic Acids Res. 21:5803-5816.) The RRM is about 80 amino acids in length and forms four β-strands and two α-helices

arranged in an α/β sandwich. The RRM contains a core RNP-1 octapeptide motif along with surrounding conserved sequences.

RNA Stability and Degradation

RNA helicases alter and regulate RNA conformation and secondary structure by using energy derived from ATP hydrolysis to destabilize and unwind RNA duplexes. The most well-characterized and ubiquitous family of RNA helicases is the DEAD-box family, so named for the conserved B-type ATP-binding motif which is diagnostic of proteins in this family. Over 40 DEAD-box helicases have been identified in organisms as diverse as bacteria, insects, yeast, amphibians, mammals, and plants. DEAD-box helicases function in diverse processes such as translation initiation, splicing, ribosome assembly, and RNA editing, transport, and stability. Some DEAD-box helicases play tissue- and stage-specific roles in spermatogenesis and embryogenesis. (Reviewed in Linder, P. et al. (1989) Nature 337:121-122.)

Overexpression of the DEAD-box 1 protein (DDX1) may play a role in the progression of neuroblastoma (Nb) and retinoblastoma (Rb) tumors. Other DEAD-box helicases have been implicated either directly or indirectly in ultraviolet light-induced tumors, B cell lymphoma, and myeloid malignancies. (Reviewed in Godbout, R. et al. (1998) J. Biol. Chem. 273:21161-21168.)

Ribonucleases (RNases) catalyze the hydrolysis of phosphodiester bonds in RNA chains, thus cleaving the RNA. For example, RNase P is a ribonucleoprotein enzyme which cleaves the 5' end of pre-tRNAs as part of their maturation process. RNase H digests the RNA strand of an RNA/DNA hybrid. Such hybrids occur in cells invaded by retroviruses, and RNase H is an important enzyme in the retroviral replication cycle. RNase H domains are often found as a domain associated with reverse transcriptases. RNase activity in serum and cell extracts is elevated in a variety of cancers and infectious diseases (Schein, C.H. (1997) Nat. Biotechnol. 15:529-536). Regulation of RNase activity is being investigated as a means to control tumor angiogenesis, allergic reactions, viral infection and replication, and fungal infections.

Protein Translation

The eukaryotic ribosome is composed of a 60S (large) subunit and a 40S (small) subunit, which together form the 80S ribosome. In addition to the 18S, 28S, 5S, and 5.8S rRNAs, the ribosome also contains more than fifty proteins. The ribosomal proteins have a prefix which denotes the subunit to which they belong, either L (large) or S (small). Three important sites are identified on the ribosome. The aminoacyl-tRNA site (A site) is where charged tRNAs (with the exception of the initiator-tRNA) bind on arrival at the ribosome. The peptidyl-tRNA site (P site) is where new peptide bonds are formed, as well as where the initiator tRNA binds. The exit site (E site) is where deacylated tRNAs bind prior to their release from the ribosome. (Translation is reviewed in Stryer, L. (1995) Biochemistry, W.H. Freeman and Company, New York NY, pp. 875-908; and Lodish, H. et al. (1995)

Molecular Cell Biology, Scientific American Books, New York NY, pp. 119-138.)

tRNA Charging

Protein biosynthesis depends on each amino acid forming a linkage with the appropriate tRNA. The aminoacyl-tRNA synthetases are responsible for the activation and correct attachment of an amino acid with its cognate tRNA. The 20 aminoacyl-tRNA synthetase enzymes can be divided into two structural classes, Class I and Class II. Autoantibodies against aminoacyl-tRNAs are generated by patients with dermatomyositis and polymyositis, and correlate strongly with complicating interstitial lung disease (ILD). These antibodies appear to be generated in response to viral infection, and coxsackie virus has been used to induce experimental viral myositis in animals.

Translation Initiation

Initiation of translation can be divided into three stages. The first stage brings an initiator transfer RNA (Met-tRNA$_i$) together with the 40S ribosomal subunit to form the 43S preinitiation complex. The second stage binds the 43S preinitiation complex to the mRNA, followed by migration of the complex to the correct AUG initiation codon. The third stage brings the 60S ribosomal subunit to the 40S subunit to generate an 80S ribosome at the initiation codon. Regulation of translation primarily involves the first and second stage in the initiation process (Pain, V.M. (1996) Eur. J. Biochem. 236:747-771).

Several initiation factors, many of which contain multiple subunits, are involved in bringing an initiator tRNA and 40S ribosomal subunit together. eIF2, a guanine nucleotide binding protein, recruits the initiator tRNA to the 40S ribosomal subunit. Only when eIF2 is bound to GTP does it associate with the initiator tRNA. eIF2B, a guanine nucleotide exchange protein, is responsible for converting eIF2 from the GDP-bound inactive form to the GTP-bound active form. Two other factors, eIF1A and eIF3 bind and stabilize the 40S subunit by interacting with 18S ribosomal RNA and specific ribosomal structural proteins. eIF3 is also involved in association of the 40S ribosomal subunit with mRNA. The Met-tRNA$_i$, eIF1A, eIF3, and 40S ribosomal subunit together make up the 43S preinitiation complex (Pain, supra).

Additional factors are required for binding of the 43S preinitiation complex to an mRNA molecule, and the process is regulated at several levels. eIF4F is a complex consisting of three proteins: eIF4E, eIF4A, and eIF4G. eIF4E recognizes and binds to the mRNA 5'-terminal m$^7$GTP cap, eIF4A is a bidirectional RNA-dependent helicase, and eIF4G is a scaffolding polypeptide. eIF4G has three binding domains. The N-terminal third of eIF4G interacts with eIF4E, the central third interacts with eIF4A, and the C-terminal third interacts with eIF3 bound to the 43S preinitiation complex. Thus, eIF4G acts as a bridge between the 40S ribosomal subunit and the mRNA (Hentze, M.W. (1997) Science 275:500-501).

The ability of eIF4F to initiate binding of the 43S preinitiation complex is regulated by

structural features of the mRNA. The mRNA molecule has an untranslated region (UTR) between the 5' cap and the AUG start codon. In some mRNAs this region forms secondary structures that impede binding of the 43S preinitiation complex. The helicase activity of eIF4A is thought to function in removing this secondary structure to facilitate binding of the 43S preinitiation complex (Pain, supra).

Translation Elongation

Elongation is the process whereby additional amino acids are joined to the initiator methionine to form the complete polypeptide chain. The elongation factors EF1α, EF1β γ, and EF2 are involved in elongating the polypeptide chain following initiation. EF1α is a GTP-binding protein. In EF1α's GTP-bound form, it brings an aminoacyl-tRNA to the ribosome's A site. The amino acid attached to the newly arrived aminoacyl-tRNA forms a peptide bond with the initiator methionine. The GTP on EF1α is hydrolyzed to GDP, and EF1α-GDP dissociates from the ribosome. EF1β γ binds EF1α -GDP and induces the dissociation of GDP from EF1α, allowing EF1α to bind GTP and a new cycle to begin.

As subsequent aminoacyl-tRNAs are brought to the ribosome, EF-G, another GTP-binding protein, catalyzes the translocation of tRNAs from the A site to the P site and finally to the E site of the ribosome. This allows the processivity of translation.

Translation Termination

The release factor eRF carries out termination of translation. eRF recognizes stop codons in the mRNA, leading to the release of the polypeptide chain from the ribosome.

Post-Translational Pathways

Proteins may be modified after translation by the addition of phosphate, sugar, prenyl, fatty acid, and other chemical groups. These modifications are often required for proper protein activity. Enzymes involved in post-translational modification include kinases, phosphatases, glycosyltransferases, and prenyltransferases. The conformation of proteins may also be modified after translation by the introduction and rearrangement of disulfide bonds (rearrangement catalyzed by protein disulfide isomerase), the isomerization of proline sidechains by prolyl isomerase, and by interactions with molecular chaperone proteins.

Proteins may also be cleaved by proteases. Such cleavage may result in activation, inactivation, or complete degradation of the protein. Proteases include serine proteases, cysteine proteases, aspartic proteases, and metalloproteases. Signal peptidase in the endoplasmic reticulum (ER) lumen cleaves the signal peptide from membrane or secretory proteins that are imported into the ER. Ubiquitin proteases are associated with the ubiquitin conjugation system (UCS), a major pathway for the degradation of cellular proteins in eukaryotic cells and some bacteria. The UCS mediates the elimination of abnormal proteins and regulates the half-lives of important regulatory proteins that control cellular processes such as gene transcription and cell cycle progression. In the UCS pathway, proteins targeted for degradation are conjugated to a ubiquitin, a small heat stable

protein. Proteins involved in the UCS include ubiquitin-activating enzyme, ubiquitin-conjugating enzymes, ubiquitin-ligases, and ubiquitin C-terminal hydrolases. The ubiquitinated protein is then recognized and degraded by the proteasome, a large, multisubunit proteolytic enzyme complex, and ubiquitin is released for reutilization by ubiquitin protease.

Lipid Metabolism

Lipids are water-insoluble, oily or greasy substances that are soluble in nonpolar solvents such as chloroform or ether. Neutral fats (triacylglycerols) serve as major fuels and energy stores. Polar lipids, such as phospholipids, sphingolipids, glycolipids, and cholesterol, are key structural components of cell membranes.

Lipid metabolism is involved in human diseases and disorders. In the arterial disease atherosclerosis, fatty lesions form on the inside of the arterial wall. These lesions promote the loss of arterial flexibility and the formation of blood clots (Guyton, A.C. Textbook of Medical Physiology (1991) W.B. Saunders Company, Philadelphia PA, pp.760-763). In Tay-Sachs disease, the $GM_2$ ganglioside (a sphingolipid) accumulates in lysosomes of the central nervous system due to a lack of the enzyme N-acetylhexosaminidase. Patients suffer nervous system degeneration leading to early death (Fauci, A.S. et al. (1998) Harrison's Principles of Internal Medicine McGraw-Hill, New York NY, p. 2171). The Niemann-Pick diseases are caused by defects in lipid metabolism. Niemann-Pick diseases types A and B are caused by accumulation of sphingomyelin (a sphingolipid) and other lipids in the central nervous system due to a defect in the enzyme sphingomyelinase, leading to neurodegeneration and lung disease. Niemann-Pick disease type C results from a defect in cholesterol transport, leading to the accumulation of sphingomyelin and cholesterol in lysosomes and a secondary reduction in sphingomyelinase activity. Neurological symptoms such as grand mal seizures, ataxia, and loss of previously learned speech, manifest 1-2 years after birth. A mutation in the NPC protein, which contains a putative cholesterol-sensing domain, was found in a mouse model of Niemann-Pick disease type C (Fauci, supra, p. 2175; Loftus, S.K. et al. (1997) Science 277:232-235). (Lipid metabolism is reviewed in Stryer, L. (1995) Biochemistry, W.H. Freeman and Company, New York NY; Lehninger, A. (1982) Principles of Biochemistry Worth Publishers, Inc., New York NY; and ExPASy "Biochemical Pathways" index of Boehringer Mannheim World Wide Web site.)

Fatty Acid Synthesis

Fatty acids are long-chain organic acids with a single carboxyl group and a long non-polar hydrocarbon tail. Long-chain fatty acids are essential components of glycolipids, phospholipids, and cholesterol, which are building blocks for biological membranes, and of triglycerides, which are biological fuel molecules. Long-chain fatty acids are also substrates for eicosanoid production, and are important in the functional modification of certain complex carbohydrates and proteins. 16-carbon and 18-carbon fatty acids are the most common.

Fatty acid synthesis occurs in the cytoplasm. In the first step, acetyl-Coenzyme A (CoA) carboxylase (ACC) synthesizes malonyl-CoA from acetyl-CoA and bicarbonate. The enzymes which catalyze the remaining reactions are covalently linked into a single polypeptide chain, referred to as the multifunctional enzyme fatty acid synthase (FAS). FAS catalyzes the synthesis of palmitate from acetyl-CoA and malonyl-CoA. FAS contains acetyl transferase, malonyl transferase, β-ketoacetyl synthase, acyl carrier protein, β-ketoacyl reductase, dehydratase, enoyl reductase, and thioesterase activities. The final product of the FAS reaction is the 16-carbon fatty acid palmitate. Further elongation, as well as unsaturation, of palmitate by accessory enzymes of the ER produces the variety of long chain fatty acids required by the individual cell. These enzymes include a NADH-cytochrome $b_5$ reductase, cytochrome $b_5$, and a desaturase.

Phospholipid and Triacylglycerol Synthesis

Triacylglycerols, also known as triglycerides and neutral fats, are major energy stores in animals. Triacylglycerols are esters of glycerol with three fatty acid chains. Glycerol-3-phosphate is produced from dihydroxyacetone phosphate by the enzyme glycerol phosphate dehydrogenase or from glycerol by glycerol kinase. Fatty acid-CoA's are produced from fatty acids by fatty acyl-CoA synthetases. Glyercol-3-phosphate is acylated with two fatty acyl-CoA's by the enzyme glycerol phosphate acyltransferase to give phosphatidate. Phosphatidate phosphatase converts phosphatidate to diacylglycerol, which is subsequently acylated to a triacylglyercol by the enzyme diglyceride acyltransferase. Phosphatidate phosphatase and diglyceride acyltransferase form a triacylglyerol synthetase complex bound to the ER membrane.

A major class of phospholipids are the phosphoglycerides, which are composed of a glycerol backbone, two fatty acid chains, and a phosphorylated alcohol. Phosphoglycerides are components of cell membranes. Principal phosphoglycerides are phosphatidyl choline, phosphatidyl ethanolamine, phosphatidyl serine, phosphatidyl inositol, and diphosphatidyl glycerol. Many enzymes involved in phosphoglyceride synthesis are associated with membranes (Meyers, R.A. (1995) Molecular Biology and Biotechnology, VCH Publishers Inc., New York NY, pp. 494-501). Phosphatidate is converted to CDP-diacylglycerol by the enzyme phosphatidate cytidylyltransferase (ExPASy ENZYME EC 2.7.7.41). Transfer of the diacylglycerol group from CDP-diacylglycerol to serine to yield phosphatidyl serine, or to inositol to yield phosphatidyl inositol, is catalyzed by the enzymes CDP-diacylglycerol-serine O-phosphatidyltransferase and CDP-diacylglycerol-inositol 3-phosphatidyltransferase, respectively (ExPASy ENZYME EC 2.7.8.8; ExPASy ENZYME EC 2.7.8.11). The enzyme phosphatidyl serine decarboxylase catalyzes the conversion of phosphatidyl serine to phosphatidyl ethanolamine, using a pyruvate cofactor (Voelker, D.R. (1997) Biochim. Biophys. Acta 1348:236-244). Phosphatidyl choline is formed using diet-derived choline by the reaction of CDP-choline with 1,2-diacylglycerol, catalyzed by diacylglycerol cholinephosphotransferase (ExPASy ENZYME 2.7.8.2).

Sterol, Steroid, and Isoprenoid Metabolism

Cholesterol, composed of four fused hydrocarbon rings with an alcohol at one end, moderates the fluidity of membranes in which it is incorporated. In addition, cholesterol is used in the synthesis of steroid hormones such as cortisol, progesterone, estrogen, and testosterone. Bile salts derived from cholesterol facilitate the digestion of lipids. Cholesterol in the skin forms a barrier that prevents excess water evaporation from the body. Farnesyl and geranylgeranyl groups, which are derived from cholesterol biosynthesis intermediates, are post-translationally added to signal transduction proteins such as ras and protein-targeting proteins such as rab. These modifications are important for the activities of these proteins (Guyton, supra; Stryer, supra, pp. 279-280, 691-702, 934).

Mammals obtain cholesterol derived from both de novo biosynthesis and the diet. The liver is the major site of cholesterol biosynthesis in mammals. Two acetyl-CoA molecules initially condense to form acetoacetyl-CoA, catalyzed by a thiolase. Acetoacetyl-CoA condenses with a third acetyl-CoA to form hydroxymethylglutaryl-CoA (HMG-CoA), catalyzed by HMG-CoA synthase. Conversion of HMG-CoA to cholesterol is accomplished via a series of enzymatic steps known as the mevalonate pathway. The rate-limiting step is the conversion of HMG-CoA to mevalonate by HMG-CoA reductase. The drug lovastatin, a potent inhibitor of HMG-CoA reductase, is given to patients to reduce their serum cholesterol levels. Other mevalonate pathway enzymes include mevalonate kinase, phosphomevalonate kinase, diphosphomevalonate decarboxylase, isopentenyldiphosphate isomerase, dimethylallyl transferase, geranyl transferase, farnesyl-diphosphate farnesyltransferase, squalene monooxygenase, lanosterol synthase, lathosterol oxidase, and 7-dehydrocholesterol reductase.

Cholesterol is used in the synthesis of steroid hormones such as cortisol, progesterone, aldosterone, estrogen, and testosterone. First, cholesterol is converted to pregnenolone by cholesterol monooxygenases. The other steroid hormones are synthesized from pregnenolone by a series of enzyme-catalyzed reactions including oxidations, isomerizations, hydroxylations, reductions, and demethylations. Examples of these enzymes include steroid Δ-isomerase, 3β-hydroxy-$\Delta^5$-steroid dehydrogenase, steroid 21-monooxygenase, steroid 19-hydroxylase, and 3β-hydroxysteroid dehydrogenase. Cholesterol is also the precursor to vitamin D.

Numerous compounds contain 5-carbon isoprene units derived from the mevalonate pathway intermediate isopentenyl pyrophosphate. Isoprenoid groups are found in vitamin K, ubiquinone, retinal, dolichol phosphate (a carrier of oligosaccharides needed for N-linked glycosylation), and farnesyl and geranylgeranyl groups that modify proteins. Enzymes involved include farnesyl transferase, polyprenyl transferases, dolichyl phosphatase, and dolichyl kinase.

Sphingolipid Metabolism

Sphingolipids are an important class of membrane lipids that contain sphingosine, a long chain amino alcohol. They are composed of one long-chain fatty acid, one polar head alcohol, and

sphingosine or sphingosine derivative. The three classes of sphingolipids are sphingomyelins, cerebrosides, and gangliosides. Sphingomyelins, which contain phosphocholine or phosphoethanolamine as their head group, are abundant in the myelin sheath surrounding nerve cells. Galactocerebrosides, which contain a glucose or galactose head group, are characteristic of the brain. Other cerebrosides are found in nonneural tissues. Gangliosides, whose head groups contain multiple sugar units, are abundant in the brain, but are also found in nonneural tissues.

Sphingolipids are built on a sphingosine backbone. Sphingosine is acylated to ceramide by the enzyme sphingosine acetyltransferase. Ceramide and phosphatidyl choline are converted to sphingomyelin by the enzyme ceramide choline phosphotransferase. Cerebrosides are synthesized by the linkage of glucose or galactose to ceramide by a transferase. Sequential addition of sugar residues to ceramide by transferase enzymes yields gangliosides.

Eicosanoid Metabolism

Eicosanoids, including prostaglandins, prostacyclin, thromboxanes, and leukotrienes, are 20-carbon molecules derived from fatty acids. Eicosanoids are signaling molecules which have roles in pain, fever, and inflammation. The precursor of all eicosanoids is arachidonate, which is generated from phospholipids by phospholipase $A_2$ and from diacylglycerols by diacylglycerol lipase. Leukotrienes are produced from arachidonate by the action of lipoxygenases. Prostaglandin synthase, reductases, and isomerases are responsible for the synthesis of the prostaglandins. Prostaglandins have roles in inflammation, blood flow, ion transport, synaptic transmission, and sleep. Prostacyclin and the thromboxanes are derived from a precursor prostaglandin by the action of prostacyclin synthase and thromboxane synthases, respectively.

Ketone Body Metabolism

Pairs of acetyl-CoA molecules derived from fatty acid oxidation in the liver can condense to form acetoacetyl-CoA, which subsequently forms acetoacetate, D-3-hydroxybutyrate, and acetone. These three products are known as ketone bodies. Enzymes involved in ketone body metabolism include HMG-CoA synthetase, HMG-CoA cleavage enzyme, D-3-hydroxybutyrate dehydrogenase, acetoacetate decarboxylase, and 3-ketoacyl-CoA transferase. Ketone bodies are a normal fuel supply of the heart and renal cortex. Acetoacetate produced by the liver is transported to cells where the acetoacetate is converted back to acetyl-CoA and enters the citric acid cycle. In times of starvation, ketone bodies produced from stored triacylglyerols become an important fuel source, especially for the brain. Abnormally high levels of ketone bodies are observed in diabetics. Diabetic coma can result if ketone body levels become too great.

Lipid Mobilization

Within cells, fatty acids are transported by cytoplasmic fatty acid binding proteins (Online Mendelian Inheritance in Man (OMIM) *134650 Fatty Acid-Binding Protein 1, Liver; FABP1).

Diazepam binding inhibitor (DBI), also known as endozepine and acyl CoA-binding protein, is an endogenous γ-aminobutyric acid (GABA) receptor ligand which is thought to down-regulate the effects of GABA. DBI binds medium- and long-chain acyl-CoA esters with very high affinity and may function as an intracellular carrier of acyl-CoA esters (OMIM *125950 Diazepam Binding Inhibitor; DBI; PROSITE PDOC00686 Acyl-CoA-binding protein signature).

Fat stored in liver and adipose triglycerides may be released by hydrolysis and transported in the blood. Free fatty acids are transported in the blood by albumin. Triacylglycerols and cholesterol esters in the blood are transported in lipoprotein particles. The particles consist of a core of hydrophobic lipids surrounded by a shell of polar lipids and apolipoproteins. The protein components serve in the solubilization of hydrophobic lipids and also contain cell-targeting signals. Lipoproteins include chylomicrons, chylomicron remnants, very-low-density lipoproteins (VLDL), intermediate-density lipoproteins (IDL), low-density lipoproteins (LDL), and high-density lipoproteins (HDL). There is a strong inverse correlation between the levels of plasma HDL and risk of premature coronary heart disease.

Triacylglycerols in chylomicrons and VLDL are hydrolyzed by lipoprotein lipases that line blood vessels in muscle and other tissues that use fatty acids. Cell surface LDL receptors bind LDL particles which are then internalized by endocytosis. Absence of the LDL receptor, the cause of the disease familial hypercholesterolemia, leads to increased plasma cholesterol levels and ultimately to atherosclerosis. Plasma cholesteryl ester transfer protein mediates the transfer of cholesteryl esters from HDL to apolipoprotein B-containing lipoproteins. Cholesteryl ester transfer protein is important in the reverse cholesterol transport system and may play a role in atherosclerosis (Yamashita, S. et al. (1997) Curr. Opin. Lipidol. 8:101-110). Macrophage scavenger receptors, which bind and internalize modified lipoproteins, play a role in lipid transport and may contribute to atherosclerosis (Greaves, D.R. et al. (1998) Curr. Opin. Lipidol. 9:425-432).

Proteins involved in cholesterol uptake and biosynthesis are tightly regulated in response to cellular cholesterol levels. The sterol regulatory element binding protein (SREBP) is a sterol-responsive transcription factor. Under normal cholesterol conditions, SREBP resides in the ER membrane. When cholesterol levels are low, a regulated cleavage of SREBP occurs which releases the extracellular domain of the protein. This cleaved domain is then transported to the nucleus where it activates the transcription of the LDL receptor gene, and genes encoding enzymes of cholesterol synthesis, by binding the sterol regulatory element (SRE) upstream of the genes (Yang, J. et al. (1995) J. Biol. Chem. 270:12152-12161). Regulation of cholesterol uptake and biosynthesis also occurs via the oxysterol-binding protein (OSBP). OSBP is a high-affinity intracellular receptor for a variety of oxysterols that down-regulate cholesterol synthesis and stimulate cholesterol esterification (Lagace, T.A. et al. (1997) Biochem. J. 326:205-213).

Beta-oxidation

Mitochondrial and peroxisomal beta-oxidation enzymes degrade saturated and unsaturated fatty acids by sequential removal of two-carbon units from CoA-activated fatty acids. The main beta-oxidation pathway degrades both saturated and unsaturated fatty acids while the auxiliary pathway performs additional steps required for the degradation of unsaturated fatty acids.

The pathways of mitochondrial and peroxisomal beta-oxidation use similar enzymes, but have different substrate specificities and functions. Mitochondria oxidize short-, medium-, and long-chain fatty acids to produce energy for cells. Mitochondrial beta-oxidation is a major energy source for cardiac and skeletal muscle. In liver, it provides ketone bodies to the peripheral circulation when glucose levels are low as in starvation, endurance exercise, and diabetes (Eaton, S. et al. (1996) Biochem. J. 320:345-357). Peroxisomes oxidize medium-, long-, and very-long-chain fatty acids, dicarboxylic fatty acids, branched fatty acids, prostaglandins, xenobiotics, and bile acid intermediates. The chief roles of peroxisomal beta-oxidation are to shorten toxic lipophilic carboxylic acids to facilitate their excretion and to shorten very-long-chain fatty acids prior to mitochondrial beta-oxidation (Mannaerts, G.P. and P.P. van Veldhoven (1993) Biochimie 75:147-158).

Enzymes involved in beta-oxidation include acyl CoA synthetase, carnitine acyltransferase, acyl CoA dehydrogenases, enoyl CoA hydratases, L-3-hydroxyacyl CoA dehydrogenase, β-ketothiolase, 2,4-dienoyl CoA reductase, and isomerase.

Lipid Cleavage and Degradation

Triglycerides are hydrolyzed to fatty acids and glycerol by lipases. Lysophospholipases (LPLs) are widely distributed enzymes that metabolize intracellular lipids, and occur in numerous isoforms. Small isoforms, approximately 15-30 kD, function as hydrolases; large isoforms, those exceeding 60 kD, function both as hydrolases and transacylases. A particular substrate for LPLs, lysophosphatidylcholine, causes lysis of cell membranes when it is formed or imported into a cell. LPLs are regulated by lipid factors including acylcarnitine, arachidonic acid, and phosphatidic acid. These lipid factors are signaling molecules important in numerous pathways, including the inflammatory response. (Anderson, R. et al. (1994) Toxicol. Appl. Pharmacol. 125:176-183; Selle, H. et al. (1993); Eur. J. Biochem. 212:411-416.)

The secretory phospholipase $A_2$ (PLA2) superfamily comprises a number of heterogeneous enzymes whose common feature is to hydrolyze the sn-2 fatty acid acyl ester bond of phosphoglycerides. Hydrolysis of the glycerophospholipids releases free fatty acids and lysophospholipids. PLA2 activity generates precursors for the biosynthesis of biologically active lipids, hydroxy fatty acids, and platelet-activating factor. PLA2 hydrolysis of the sn-2 ester bond in phospholipids generates free fatty acids, such as arachidonic acid and lysophospholipids.

Carbon and Carbohydrate Metabolism

Carbohydrates, including sugars or saccharides, starch, and cellulose, are aldehyde or ketone compounds with multiple hydroxyl groups. The importance of carbohydrate metabolism is demonstrated by the sensitive regulatory system in place for maintenance of blood glucose levels. Two pancreatic hormones, insulin and glucagon, promote increased glucose uptake and storage by cells, and increased glucose release from cells, respectively. Carbohydrates have three important roles in mammalian cells. First, carbohydrates are used as energy stores, fuels, and metabolic intermediates. Carbohydrates are broken down to form energy in glycolysis and are stored as glycogen for later use. Second, the sugars deoxyribose and ribose form part of the structural support of DNA and RNA, respectively. Third, carbohydrate modifications are added to secreted and membrane proteins and lipids as they traverse the secretory pathway. Cell surface carbohydrate-containing macromolecules, including glycoproteins, glycolipids, and transmembrane proteoglycans, mediate adhesion with other cells and with components of the extracellular matrix. The extracellular matrix is comprised of diverse glycoproteins, glycosaminoglycans (GAGs), and carbohydrate-binding proteins which are secreted from the cell and assembled into an organized meshwork in close association with the cell surface. The interaction of the cell with the surrounding matrix profoundly influences cell shape, strength, flexibility, motility, and adhesion. These dynamic properties are intimately associated with signal transduction pathways controlling cell proliferation and differentiation, tissue construction, and embryonic development.

Carbohydrate metabolism is altered in several disorders including diabetes mellitus, hyperglycemia, hypoglycemia, galactosemia, galactokinase deficiency, and UDP-galactose-4-epimerase deficiency (Fauci, A.S. et al. (1998) Harrison's Principles of Internal Medicine, McGraw-Hill, New York NY, pp. 2208-2209). Altered carbohydrate metabolism is associated with cancer. Reduced GAG and proteoglycan expression is associated with human lung carcinomas (Nackaerts, K. et al. (1997) Int. J. Cancer 74:335-345). The carbohydrate determinants sialyl Lewis A and sialyl Lewis X are frequently expressed on human cancer cells (Kannagi, R. (1997) Glycoconj. J. 14:577-584). Alterations of the N-linked carbohydrate core structure of cell surface glycoproteins are linked to colon and pancreatic cancers (Schwarz, R.E. et al. (1996) Cancer Lett. 107:285-291). Reduced expression of the Sda blood group carbohydrate structure in cell surface glycolipids and glycoproteins is observed in gastrointestinal cancer (Dohi, T. et al. (1996) Int. J. Cancer 67:626-663). (Carbon and carbohydrate metabolism is reviewed in Stryer, L. (1995) Biochemistry W.H. Freeman and Company, New York NY; Lehninger, A.L. (1982) Principles of Biochemistry Worth Publishers Inc., New York NY; and Lodish, H. et al. (1995) Molecular Cell Biology Scientific American Books, New York NY.)

Glycolysis

Enzymes of the glycolytic pathway convert the sugar glucose to pyruvate while simultaneously producing ATP. The pathway also provides building blocks for the synthesis of cellular components

such as long-chain fatty acids. After glycolysis, pyrvuate is converted to acetyl-Coenzyme A, which, in aerobic organisms, enters the citric acid cycle. Glycolytic enzymes include hexokinase, phosphoglucose isomerase, phosphofructokinase, aldolase, triose phosphate isomerase, glyceraldehyde 3-phosphate dehydrogenase, phosphoglycerate kinase, phosphoglyceromutase, enolase, and pyruvate kinase. Of these, phosphofructokinase, hexokinase, and pyruvate kinase are important in regulating the rate of glycolysis.

Gluconeogenesis

Gluconeogenesis is the synthesis of glucose from noncarbohydrate precursors such as lactate and amino acids. The pathway, which functions mainly in times of starvation and intense exercise, occurs mostly in the liver and kidney. Responsible enzymes include pyruvate carboxylase, phosphoenolpyruvate carboxykinase, fructose 1,6-bisphosphatase, and glucose-6-phosphatase.

Pentose Phosphate Pathway

Pentose phosphate pathway enzymes are responsible for generating the reducing agent NADPH, while at the same time oxidizing glucose-6-phosphate to ribose-5-phosphate. Ribose-5-phosphate and its derivatives become part of important biological molecules such as ATP, Coenzyme A, $NAD^+$, FAD, RNA, and DNA. The pentose phosphate pathway has both oxidative and non-oxidative branches. The oxidative branch steps, which are catalyzed by the enzymes glucose-6-phosphate dehydrogenase, lactonase, and 6-phosphogluconate dehydrogenase, convert glucose-6-phosphate and $NADP^+$ to ribulose-6-phosphate and NADPH. The non-oxidative branch steps, which are catalyzed by the enzymes phosphopentose isomerase, phosphopentose epimerase, transketolase, and transaldolase, allow the interconversion of three-, four-, five-, six-, and seven-carbon sugars.

Glucouronate Metabolism

Glucuronate is a monosaccharide which, in the form of D-glucuronic acid, is found in the GAGs chondroitin and dermatan. D-glucuronic acid is also important in the detoxification and excretion of foreign organic compounds such as phenol. Enzymes involved in glucuronate metabolism include UDP-glucose dehydrogenase and glucuronate reductase.

Disaccharide Metabolism

Disaccharides must be hydrolyzed to monosaccharides to be digested. Lactose, a disaccharide found in milk, is hydrolyzed to galactose and glucose by the enzyme lactase. Maltose is derived from plant starch and is hydrolyzed to glucose by the enzyme maltase. Sucrose is derived from plants and is hydrolyzed to glucose and fructose by the enzyme sucrase. Trehalose, a disaccharide found mainly in insects and mushrooms, is hydrolyzed to glucose by the enzyme trehalase (OMIM *275360 Trehalase; Ruf, J. et al. (1990) J. Biol. Chem. 265:15034-15039). Lactase, maltase, sucrase, and trehalase are bound to mucosal cells lining the small intestine, where they participate in the digestion of dietary disaccharides. The enzyme lactose synthetase, composed of the catalytic subunit galactosyltransferase

and the modifier subunit α-lactalbumin, converts UDP-galactose and glucose to lactose in the mammary glands.

Glycogen, Starch, and Chitin Metabolism

Glycogen is the storage form of carbohydrates in mammals. Mobilization of glycogen maintains glucose levels between meals and during muscular activity. Glycogen is stored mainly in the liver and in skeletal muscle in the form of cytoplasmic granules. These granules contain enzymes that catalyze the synthesis and degradation of glycogen, as well as enzymes that regulate these processes. Enzymes that catalyze the degradation of glycogen include glycogen phosphorylase, a transferase, α-1,6-glucosidase, and phosphoglucomutase. Enzymes that catalyze the synthesis of glycogen include UDP-glucose pyrophosphorylase, glycogen synthetase, a branching enzyme, and nucleoside diphosphokinase. The enzymes of glycogen synthesis and degradation are tightly regulated by the hormones insulin, glucagon, and epinephrine. Starch, a plant-derived polysaccharide, is hydrolyzed to maltose, maltotriose, and α-dextrin by α-amylase, an enzyme secreted by the salivary glands and pancreas. Chitin is a polysaccharide found in insects and crustacea. A chitotriosidase is secreted by macrophages and may play a role in the degradation of chitin-containing pathogens (Boot, R.G. et al. (1995) J. Biol. Chem. 270:26252-26256).

Peptidoglycans and Glycosaminoglycans

Glycosaminoglycans (GAGs) are anionic linear unbranched polysaccharides composed of repetitive disaccharide units. These repetitive units contain a derivative of an amino sugar, either glucosamine or galactosamine. GAGs exist free or as part of proteoglycans, large molecules composed of a core protein attached to one or more GAGs. GAGs are found on the cell surface, inside cells, and in the extracellular matrix. Changes in GAG levels are associated with several autoimmune diseases including autoimmune thyroid disease, autoimmune diabetes mellitus, and systemic lupus erythematosus (Hansen, C. et al. (1996) Clin. Exp. Rheum. 14 (Suppl. 15):S59-S67). GAGs include chondroitin sulfate, keratan sulfate, heparin, heparan sulfate, dermatan sulfate, and hyaluronan.

The GAG hyaluronan (HA) is found in the extracellular matrix of many cells, especially in soft connective tissues, and is abundant in synovial fluid (Pitsillides, A.A. et al. (1993) Int. J. Exp. Pathol. 74:27-34). HA seems to play important roles in cell regulation, development, and differentiation (Laurent, T.C. and J.R. Fraser (1992) FASEB J. 6:2397-2404). Hyaluronidase is an enzyme that degrades HA to oligosaccharides. Hyaluronidases may function in cell adhesion, infection, angiogenesis, signal transduction, reproduction, cancer, and inflammation.

Proteoglycans, also known as peptidoglycans, are found in the extracellular matrix of connective tissues such as cartilage and are essential for distributing the load in weight-bearing joints. Cell-surface-attached proteoglycans anchor cells to the extracellular matrix. Both extracellular and cell-surface proteoglycans bind growth factors, facilitating their binding to cell-surface receptors and

subsequent triggering of signal transduction pathways.

Amino Acid and Nitrogen Metabolism

$NH_4^+$ is assimilated into amino acids by the actions of two enzymes, glutamate dehydrogenase and glutamine synthetase. The carbon skeletons of amino acids come from the intermediates of glycolysis, the pentose phosphate pathway, or the citric acid cycle. Of the twenty amino acids used in proteins, humans can synthesize only thirteen (nonessential amino acids). The remaining nine must come from the diet (essential amino acids). Enzymes involved in nonessential amino acid biosynthesis include glutamate kinase dehydrogenase, pyrroline carboxylate reductase, asparagine synthetase, phenylalanine oxygenase, methionine adenosyltransferase, adenosylhomocysteinase, cystathionine β-synthase, cystathionine γ-lyase, phosphoglycerate dehydrogenase, phosphoserine transaminase, phosphoserine phosphatase, serine hydroxylmethyltransferase, and glycine synthase.

Metabolism of amino acids takes place almost entirely in the liver, where the amino group is removed by aminotransferases (transaminases), for example, alanine aminotransferase. The amino group is transferred to α-ketoglutarate to form glutamate. Glutamate dehydrogenase converts glutamate to $NH_4^+$ and α-ketoglutarate. $NH_4^+$ is converted to urea by the urea cycle which is catalyzed by the enzymes arginase, ornithine transcarbamoylase, arginosuccinate synthetase, and arginosuccinase. Carbamoyl phosphate synthetase is also involved in urea formation. Enzymes involved in the metabolism of the carbon skeleton of amino acids include serine dehydratase, asparaginase, glutaminase, propionyl CoA carboxylase, methylmalonyl CoA mutase, branched-chain α-keto dehydrogenase complex, isovaleryl CoA dehydrogenase, β-methylcrotonyl CoA carboxylase, phenylalanine hydroxylase, p-hydroxylphenylpyruvate hydroxylase, and homogentisate oxidase.

Polyamines, which include spermidine, putrescine, and spermine, bind tightly to nucleic acids and are abundant in rapidly proliferating cells. Enzymes involved in polyamine synthesis include ornithine decarboxylase.

Diseases involved in amino acid and nitrogen metabolism include hyperammonemia, carbamoyl phosphate synthetase deficiency, urea cycle enzyme deficiencies, methylmalonic aciduria, maple syrup disease, alcaptonuria, and phenylketonuria.

Energy Metabolism

Cells derive energy from metabolism of ingested compounds that may be roughly categorized as carbohydrates, fats, or proteins. Energy is also stored in polymers such as triglycerides (fats) and glycogen (carbohydrates). Metabolism proceeds along separate reaction pathways connected by key intermediates such as acetyl coenzyme A (acetyl-CoA). Metabolic pathways feature anaerobic and aerobic degradation, coupled with the energy-requiring reactions such as phosphorylation of adenosine diphosphate (ADP) to the triphosphate (ATP) or analogous phosphorylations of guanosine

(GDP/GTP), uridine (UDP/UTP), or cytidine (CDP/CTP). Subsequent dephosphorylation of the triphosphate drives reactions needed for cell maintenance, growth, and proliferation.

Digestive enzymes convert carbohydrates and sugars to glucose; fructose and galactose are converted in the liver to glucose. Enzymes involved in these conversions include galactose-1-phosphate uridyl transferase and UDP-galactose-4 epimerase. In the cytoplasm, glycolysis converts glucose to pyruvate in a series of reactions coupled to ATP synthesis.

Pyruvate is transported into the mitochondria and converted to acetyl-CoA for oxidation via the citric acid cycle, involving pyruvate dehydrogenase components, dihydrolipoyl transacetylase, and dihydrolipoyl dehydrogenase. Enzymes involved in the citric acid cycle include: citrate synthetase, aconitases, isocitrate dehydrogenase, alpha-ketoglutarate dehydrogenase complex including transsuccinylases, succinyl CoA synthetase, succinate dehydrogenase, fumarases, and malate dehydrogenase. Acetyl CoA is oxidized to $CO_2$ with concomitant formation of NADH, $FADH_2$, and GTP. In oxidative phosphorylation, the transport of electrons from NADH and $FADH_2$ to oxygen by dehydrogenases is coupled to the synthesis of ATP from ADP and $P_i$ by the $F_0F_1$ ATPase complex in the mitochondrial inner membrane. Enzyme complexes responsible for electron transport and ATP synthesis include the $F_0F_1$ ATPase complex, ubiquinone(CoQ)-cytochrome c reductase, ubiquinone reductase, cytochrome b, cytochrome $c_1$, FeS protein, and cytochrome c oxidase.

Triglycerides are hydrolyzed to fatty acids and glycerol by lipases. Glycerol is then phosphorylated to glycerol-3-phosphate by glycerol kinase and glycerol phosphate dehydrogenase, and degraded by the glycolysis. Fatty acids are transported into the mitochondria as fatty acyl-carnitine esters and undergo oxidative degradation.

In addition to metabolic disorders such as diabetes and obesity, disorders of energy metabolism are associated with cancers (Dorward, A. et al. (1997) J. Bioenerg. Biomembr. 29:385-392), autism (Lombard, J. (1998) Med. Hypotheses 50:497-500), neurodegenerative disorders (Alexi, T. et al. (1998) Neuroreport 9:R57-64), and neuromuscular disorders (DiMauro, S. et al. (1998) Biochim. Biophys. Acta 1366:199-210). The myocardium is heavily dependent on oxidative metabolism, so metabolic dysfunction often leads to heart disease (DiMauro, S. and M. Hirano (1998) Curr. Opin. Cardiol. 13:190-197).

For a review of energy metabolism enzymes and intermediates, see Stryer, L. et al. (1995) Biochemistry, W.H. Freeman and Co., San Francisco CA, pp. 443-652. For a review of energy metabolism regulation, see Lodish, H. et al. (1995) Molecular Cell Biology, Scientific American Books, New York NY, pp. 744-770.

Cofactor Metabolism

Cofactors, including coenzymes and prosthetic groups, are small molecular weight inorganic or organic compounds that are required for the action of an enzyme. Many cofactors contain vitamins

as a component. Cofactors include thiamine pyrophosphate, flavin adenine dinucleotide, flavin mononucleotide, nicotinamide adenine dinucleotide, pyridoxal phosphate, coenzyme A, tetrahydrofolate, lipoamide, and heme. The vitamins biotin and cobalamin are associated with enzymes as well. Heme, a prosthetic group found in myoglobin and hemoglobin, consists of protoporphyrin group bound to iron. Porphyrin groups contain four substituted pyrroles covalently joined in a ring, often with a bound metal atom. Enzymes involved in porphyrin synthesis include δ-aminolevulinate synthase, δ-aminolevulinate dehydrase, porphobilinogen deaminase, and cosynthase. Deficiencies in heme formation cause porphyrias. Heme is broken down as a part of erythrocyte turnover. Enzymes involved in heme degradation include heme oxygenase and biliverdin reductase.

Iron is a required cofactor for many enzymes. Besides the heme-containing enzymes, iron is found in iron-sulfur clusters in proteins including aconitase, succinate dehydrogenase, and NADH-Q reductase. Iron is transported in the blood by the protein transferrin. Binding of transferrin to the transferrin receptor on cell surfaces allows uptake by receptor mediated endocytosis. Cytosolic iron is bound to ferritin protein.

A molybdenum-containing cofactor (molybdopterin) is found in enzymes including sulfite oxidase, xanthine dehydrogenase, and aldehyde oxidase. Molybdopterin biosynthesis is performed by two molybdenum cofactor synthesizing enzymes. Deficiencies in these enzymes cause mental retardation and lens dislocation. Other diseases caused by defects in cofactor metabolism include pernicious anemia and methylmalonic aciduria.

Secretion and Trafficking

Eukaryotic cells are bound by a lipid bilayer membrane and subdivided into functionally distinct, membrane bound compartments. The membranes maintain the essential differences between the cytosol, the extracellular environment, and the lumenal space of each intracellular organelle. As lipid membranes are highly impermeable to most polar molecules, transport of essential nutrients, metabolic waste products, cell signaling molecules, macromolecules and proteins across lipid membranes and between organelles must be mediated by a variety of transport-associated molecules.

Protein Trafficking

In eukaryotes, some proteins are synthesized on ER-bound ribosomes, co-translationally imported into the ER, delivered from the ER to the Golgi complex for post-translational processing and sorting, and transported from the Golgi to specific intracellular and extracellular destinations. All cells possess a constitutive transport process which maintains homeostasis between the cell and its environment. In many differentiated cell types, the basic machinery is modified to carry out specific transport functions. For example, in endocrine glands, hormones and other secreted proteins are packaged into secretory granules for regulated exocytosis to the cell exterior. In macrophage, foreign extracellular material is engulfed (phagocytosis) and delivered to lysosomes for degradation. In fat and

muscle cells, glucose transporters are stored in vesicles which fuse with the plasma membrane only in response to insulin stimulation.

The Secretory Pathway

Synthesis of most integral membrane proteins, secreted proteins, and proteins destined for the lumen of a particular organelle occurs on ER-bound ribosomes. These proteins are co-translationally imported into the ER. The proteins leave the ER via membrane-bound vesicles which bud off the ER at specific sites and fuse with each other (homotypic fusion) to form the ER-Golgi Intermediate Compartment (ERGIC). The ERGIC matures progressively through the *cis*, *medial*, and *trans* cisternal stacks of the Golgi, modifying the enzyme composition by retrograde transport of specific Golgi enzymes. In this way, proteins moving through the Golgi undergo post-translational modification, such as glycosylation. The final Golgi compartment is the Trans-Golgi Network (TGN), where both membrane and lumenal proteins are sorted for their final destination. Transport vesicles destined for intracellular compartments, such as the lysosome, bud off the TGN. What remains is a secretory vesicle which contains proteins destined for the plasma membrane, such as receptors, adhesion molecules, and ion channels, and secretory proteins, such as hormones, neurotransmitters, and digestive enzymes. Secretory vesicles eventually fuse with the plasma membrane (Glick, B.S. and V. Malhotra (1998) Cell 95:883-889).

The secretory process can be constitutive or regulated. Most cells have a constitutive pathway for secretion, whereby vesicles derived from maturation of the TGN require no specific signal to fuse with the plasma membrane. In many cells, such as endocrine cells, digestive cells, and neurons, vesicle pools derived from the TGN collect in the cytoplasm and do not fuse with the plasma membrane until they are directed to by a specific signal.

Endocytosis

Endocytosis, wherein cells internalize material from the extracellular environment, is essential for transmission of neuronal, metabolic, and proliferative signals; uptake of many essential nutrients; and defense against invading organisms. Most cells exhibit two forms of endocytosis. The first, phagocytosis, is an actin-driven process exemplified in macrophage and neutrophils. Material to be endocytosed contacts numerous cell surface receptors which stimulate the plasma membrane to extend and surround the particle, enclosing it in a membrane-bound phagosome. In the mammalian immune system, IgG-coated particles bind Fc receptors on the surface of phagocytic leukocytes. Activation of the Fc receptors initiates a signal cascade involving src-family cytosolic kinases and the monomeric GTP-binding (G) protein Rho. The resulting actin reorganization leads to phagocytosis of the particle. This process is an important component of the humoral immune response, allowing the processing and presentation of bacterial-derived peptides to antigen-specific T-lymphocytes.

The second form of endocytosis, pinocytosis, is a more generalized uptake of material from the

external milieu. Like phagocytosis, pinocytosis is activated by ligand binding to cell surface receptors. Activation of individual receptors stimulates an internal response that includes coalescence of the receptor-ligand complexes and formation of clathrin-coated pits. Invagination of the plasma membrane at clathrin-coated pits produces an endocytic vesicle within the cell cytoplasm. These vesicles undergo homotypic fusion to form an early endosomal (EE) compartment. The tubulovesicular EE serves as a sorting site for incoming material. ATP-driven proton pumps in the EE membrane lowers the pH of the EE lumen (pH 6.3-6.8). The acidic environment causes many ligands to dissociate from their receptors. The receptors, along with membrane and other integral membrane proteins, are recycled back to the plasma membrane by budding off the tubular extensions of the EE in recycling vesicles (RV). This selective removal of recycled components produces a carrier vesicle containing ligand and other material from the external environment. The carrier vesicle fuses with TGN-derived vesicles which contain hydrolytic enzymes. The acidic environment of the resulting late endosome (LE) activates the hydrolytic enzymes which degrade the ligands and other material. As digestion takes place, the LE fuses with the lysosome where digestion is completed (Mellman, I. (1996) Annu. Rev. Cell Dev. Biol. 12:575-625).

Recycling vesicles may return directly to the plasma membrane. Receptors internalized and returned directly to the plasma membrane have a turnover rate of 2-3 minutes. Some RVs undergo microtubule-directed relocation to a perinuclear site, from which they then return to the plasma membrane. Receptors following this route have a turnover rate of 5-10 minutes. Still other RVs are retained within the cell until an appropriate signal is received (Mellman, supra; and James, D.E. et al. (1994) Trends Cell Biol. 4:120-126).

Vesicle Formation

Several steps in the transit of material along the secretory and endocytic pathways require the formation of transport vesicles. Specifically, vesicles form at the transitional endoplasmic reticulum (tER), the rim of Golgi cisternae, the face of the Trans-Golgi Network (TGN), the plasma membrane (PM), and tubular extensions of the endosomes. The process begins with the budding of a vesicle out of the donor membrane. The membrane-bound vesicle contains proteins to be transported and is surrounded by a protective coat made up of protein subunits recruited from the cytosol. The initial budding and coating processes are controlled by a cytosolic ras-like GTP-binding protein, ADP-ribosylating factor (Arf), and adapter proteins (AP). Different isoforms of both Arf and AP are involved at different sites of budding. Another small G-protein, dynamin, forms a ring complex around the neck of the forming vesicle and may provide the mechanochemical force to accomplish the final step of the budding process. The coated vesicle complex is then transported through the cytosol. During the transport process, Arf-bound GTP is hydrolyzed to GDP and the coat dissociates from the transport vesicle (West, M.A. et al. (1997) J. Cell Biol. 138:1239-1254). Two different classes of coat protein

have also been identified. Clathrin coats form on the TGN and PM surfaces, whereas coatomer or COP coats form on the ER and Golgi. COP coats can further be distinguished as COPI, involved in retrograde traffic through the Golgi and from the Golgi to the ER, and COPII, involved in anterograde traffic from the ER to the Golgi (Mellman, supra). The COP coat consists of two major components, a G-protein (Arf or Sar) and coat protomer (coatomer). Coatomer is an equimolar complex of seven proteins, termed alpha-, beta-, beta'-, gamma-, delta-, epsilon- and zeta-COP. (Harter, C. and F.T. Wieland (1998) Proc. Natl. Acad. Sci. USA 95:11649-11654.)

Membrane Fusion

Transport vesicles undergo homotypic or heterotypic fusion in the secretory and endocytotic pathways. Molecules required for appropriate targeting and fusion of vesicles with their target membrane include proteins incorporated in the vesicle membrane, the target membrane, and proteins recruited from the cytosol. During budding of the vesicle from the donor compartment, an integral membrane protein, VAMP (vesicle-associated membrane protein) is incorporated into the vesicle. Soon after the vesicle uncoats, a cytosolic prenylated GTP-binding protein, Rab (a member of the Ras superfamily), is inserted into the vesicle membrane. GTP-bound Rab proteins are directed into nascent transport vesicles where they interact with VAMP. Following vesicle transport, GTPase activating proteins (GAPs) in the target membrane convert Rab proteins to the GDP-bound form. A cytosolic protein, guanine-nucleotide dissociation inhibitor (GDI) helps return GDP-bound Rab proteins to their membrane of origin. Several Rab isoforms have been identified and appear to associate with specific compartments within the cell. Rab proteins appear to play a role in mediating the function of a viral gene, Rev, which is essential for replication of HIV-1, the virus responsible for AIDS (Flavell, R.A. et al. (1996) Proc. Natl. Acad. Sci. USA 93:4421-4424).

Docking of the transport vesicle with the target membrane involves the formation of a complex between the vesicle SNAP receptor (v-SNARE), target membrane (t-) SNAREs, and certain other membrane and cytosolic proteins. Many of these other proteins have been identified although their exact functions in the docking complex remain uncertain (Tellam, J.T. et al. (1995) J. Biol. Chem. 270:5857-63; and Hata, Y. and T.C. Sudhof (1995) J. Biol. Chem. 270:13022-28). N-ethylmaleimide sensitive factor (NSF) and soluble NSF-attachment protein (α-SNAP and β-SNAP) are two such proteins that are conserved from yeast to man and function in most intracellular membrane fusion reactions. Sec1 represents a family of yeast proteins that function at many different stages in the secretory pathway including membrane fusion. Recently, mammalian homologs of Sec1, called Munc-18 proteins, have been identified (Katagiri, H. et al. (1995) J. Biol. Chem. 270:4963-4966; Hata et al. supra).

The SNARE complex involves three SNARE molecules, one in the vesicular membrane and two in the target membrane. Synaptotagmin is an integral membrane protein in the synaptic vesicle

which associates with the t-SNARE syntaxin in the docking complex. Synaptotagmin binds calcium in a complex with negatively charged phospholipids, which allows the cytosolic SNAP protein to displace synaptotagmin from syntaxin and fusion to occur. Thus, synaptotagmin is a negative regulator of fusion in the neuron (Littleton, J.T. et al. (1993) Cell 74:1125-1134). The most abundant membrane protein of synaptic vesicles appears to be the glycoprotein synaptophysin, a 38 kDa protein with four transmembrane domains.

Specificity between a vesicle and its target is derived from the v-SNARE, t-SNAREs, and associated proteins involved. Different isoforms of SNAREs and Rabs show distinct cellular and subcellular distributions. VAMP-1/synaptobrevin, membrane-anchored synaptosome-associated protein of 25 kDa (SNAP-25), syntaxin-1, Rab3A, Rab15, and Rab23 are predominantly expressed in the brain and nervous system. Different syntaxin, VAMP, and Rab proteins are associated with distinct subcellular compartments and their vesicular carriers.

Nuclear Transport

Transport of proteins and RNA between the nucleus and the cytoplasm occurs through nuclear pore complexes (NPCs). NPC-mediated transport occurs in both directions through the nuclear envelope. All nuclear proteins are imported from the cytoplasm, their site of synthesis. tRNA and mRNA are exported from the nucleus, their site of synthesis, to the cytoplasm, their site of function. Processing of small nuclear RNAs involves export into the cytoplasm, assembly with proteins and modifications such as hypermethylation to produce small nuclear ribonuclear proteins (snRNPs), and subsequent import of the snRNPs back into the nucleus. The assembly of ribosomes requires the initial import of ribosomal proteins from the cytoplasm, their incorporation with RNA into ribosomal subunits, and export back to the cytoplasm. (Görlich, D. and I.W. Mattaj (1996) Science 271:1513-1518.)

The transport of proteins and mRNAs across the NPC is selective, dependent on nuclear localization signals, and generally requires association with nuclear transport factors. Nuclear localization signals (NLS) consist of short stretches of amino acids enriched in basic residues. NLS are found on proteins that are targeted to the nucleus, such as the glucocorticoid receptor. The NLS is recognized by the NLS receptor, importin, which then interacts with the monomeric GTP-binding protein Ran. This NLS protein/receptor/Ran complex navigates the nuclear pore with the help of the homodimeric protein nuclear transport factor 2 (NTF2). NTF2 binds the GDP-bound form of Ran and to multiple proteins of the nuclear pore complex containing FXFG repeat motifs, such as p62. (Paschal, B. et al. (1997) J. Biol. Chem. 272:21534-21539; and Wong, D.H. et al. (1997) Mol. Cell Biol. 17:3755-3767). Some proteins are dissociated before nuclear mRNAs are transported across the NPC while others are dissociated shortly after nuclear mRNA transport across the NPC and are reimported into the nucleus.

Disease Correlation

The etiology of numerous human diseases and disorders can be attributed to defects in the transport or secretion of proteins. For example, abnormal hormonal secretion is linked to disorders such as diabetes insipidus (vasopressin), hyper- and hypoglycemia (insulin, glucagon), Grave's disease and goiter (thyroid hormone), and Cushing's and Addison's diseases (adrenocorticotropic hormone, ACTH). Moreover, cancer cells secrete excessive amounts of hormones or other biologically active peptides. Disorders related to excessive secretion of biologically active peptides by tumor cells include fasting hypoglycemia due to increased insulin secretion from insulinoma-islet cell tumors; hypertension due to increased epinephrine and norepinephrine secreted from pheochromocytomas of the adrenal medulla and sympathetic paraganglia; and carcinoid syndrome, which is characterized by abdominal cramps, diarrhea, and valvular heart disease caused by excessive amounts of vasoactive substances such as serotonin, bradykinin, histamine, prostaglandins, and polypeptide hormones, secreted from intestinal tumors. Biologically active peptides that are ectopically synthesized in and secreted from tumor cells include ACTH and vasopressin (lung and pancreatic cancers); parathyroid hormone (lung and bladder cancers); calcitonin (lung and breast cancers); and thyroid-stimulating hormone (medullary thyroid carcinoma). Such peptides may be useful as diagnostic markers for tumorigenesis (Schwartz, M.Z. (1997) Semin. Pediatr. Surg. 3:141-146; and Said, S.I. and G.R. Faloona (1975) N. Engl. J. Med. 293:155-160).

Defective nuclear transport may play a role in cancer. The BRCA1 protein contains three potential NLSs which interact with importin alpha, and is transported into the nucleus by the importin/NPC pathway. In breast cancer cells the BRCA1 protein is aberrantly localized in the cytoplasm. The mislocation of the BRCA1 protein in breast cancer cells may be due to a defect in the NPC nuclear import pathway (Chen, C.F. et al. (1996) J. Biol. Chem. 271:32863-32868).

It has been suggested that in some breast cancers, the tumor-suppressing activity of p53 is inactivated by the sequestration of the protein in the cytoplasm, away from its site of action in the cell nucleus. Cytoplasmic wild-type p53 was also found in human cervical carcinoma cell lines. (Moll, U.M. et al. (1992) Proc. Natl. Acad. Sci. USA 89:7262-7266; and Liang, X.H. et al. (1993) Oncogene 8:2645-2652.)

Environmental Responses

Organisms respond to the environment by a number of pathways. Heat shock proteins, including hsp 70, hsp60, hsp90, and hsp 40, assist organisms in coping with heat damage to cellular proteins.

Aquaporins (AQP) are channels that transport water and, in some cases, nonionic small solutes such as urea and glycerol. Water movement is important for a number of physiological processes including renal fluid filtration, aqueous humor generation in the eye, cerebrospinal fluid production in

the brain, and appropriate hydration of the lung. Aquaporins are members of the major intrinsic protein (MIP) family of membrane transporters (King, L.S. and P. Agre (1996) Annu. Rev. Physiol. 58:619-648; Ishibashi, K. et al. (1997) J. Biol. Chem. 272:20782-20786). The study of aquaporins may have relevance to understanding edema formation and fluid balance in both normal physiology and disease states (King, supra). Mutations in AQP2 cause autosomal recessive nephrogenic diabetes insipidus (OMIM *107777 Aquaporin 2; AQP2). Reduced AQP4 expression in skeletal muscle may be associated with Duchenne muscular dystrophy (Frigeri, A. et al. (1998) J. Clin. Invest. 102:695-703). Mutations in AQP0 cause autosomal dominant cataracts in the mouse (OMIM *154050 Major Intrinsic Protein of Lens Fiber; MIP).

The metallothioneins (MTs) are a group of small (61 amino acids), cysteine-rich proteins that bind heavy metals such as cadmium, zinc, mercury, lead, and copper and are thought to play a role in metal detoxification or the metabolism and homeostasis of metals. Arsenite-resistance proteins have been identified in hamsters that are resistant to toxic levels of arsenite (Rossman, T.G. et al. (1997) Mutat. Res. 386:307-314).

Humans respond to light and odors by specific protein pathways. Proteins involved in light perception include rhodopsin, transducin, and cGMP phosphodiesterase. Proteins involved in odor perception include multiple olfactory receptors. Other proteins are important in human Circadian rhythms and responses to wounds.

Immunity and Host Defense

All vertebrates have developed sophisticated and complex immune systems that provide protection from viral, bacterial, fungal and parasitic infections. Included in these systems are the processes of humoral immunity, the complement cascade and the inflammatory response (Paul, W.E. (1993) Fundamental Immunology, Raven Press, Ltd., New York NY, pp.1-20).

The cellular components of the humoral immune system include six different types of leukocytes: monocytes, lymphocytes, polymorphonuclear granulocytes (consisting of neutrophils, eosinophils, and basophils) and plasma cells. Additionally, fragments of megakaryocytes, a seventh type of white blood cell in the bone marrow, occur in large numbers in the blood as platelets.

Leukocytes are formed from two stem cell lineages in bone marrow. The myeloid stem cell line produces granulocytes and monocytes and, the lymphoid stem cell produces lymphocytes. Lymphoid cells travel to the thymus, spleen and lymph nodes, where they mature and differentiate into lymphocytes. Leukocytes are responsible for defending the body against invading pathogens. Neutrophils and monocytes attack invading bacteria, viruses, and other pathogens and destroy them by phagocytosis. Monocytes enter tissues and differentiate into macrophages which are extremely phagocytic. Lymphocytes and plasma cells are a part of the immune system which recognizes specific foreign molecules and organisms and inactivates them, as well as signals other cells to attack

the invaders.

Granulocytes and monocytes are formed and stored in the bone marrow until needed. Megakaryocytes are produced in bone marrow, where they fragment into platelets and are released into the bloodstream. The main function of platelets is to activate the blood clotting mechanism. Lymphocytes and plasma cells are produced in various lymphogenous organs, including the lymph nodes, spleen, thymus, and tonsils.

Both neutrophils and macrophages exhibit chemotaxis towards sites of inflammation. Tissue inflammation in response to pathogen invasion results in production of chemo-attractants for leukocytes, such as endotoxins or other bacterial products, prostaglandins, and products of leukocytes or platelets.

Basophils participate in the release of the chemicals involved in the inflammatory process. The main function of basophils is secretion of these chemicals to such a degree that they have been referred to as "unicellular endocrine glands". A distinct aspect of basophilic secretion is that the contents of granules go directly into the extracellular environment, not into vacuoles as occurs with neutrophils, eosinophils and monocytes. Basophils have receptors for the Fc fragment of immunoglobulin E (IgE) that are not present on other leukocytes. Crosslinking of membrane IgE with anti-IgE or other ligands triggers degranulation.

Eosinophils are bi- or multi-nucleated white blood cells which contain eosinophilic granules. Their plasma membrane is characterized by Ig receptors, particularly IgG and IgE. Generally, eosinophils are stored in the bone marrow until recruited for use at a site of inflammation or invasion. They have specific functions in parasitic infections and allergic reactions, and are thought to detoxify some of the substances released by mast cells and basophils which cause inflammation. Additionally, they phagocytize antigen-antibody complexes and further help prevent spread of the inflammation.

Macrophages are monocytes that have left the blood stream to settle in tissue. Once monocytes have migrated into tissues, they do not re-enter the bloodstream. The mononuclear phagocyte system is comprised of precursor cells in the bone marrow, monocytes in circulation, and macrophages in tissues. The system is capable of very fast and extensive phagocytosis. A macrophage may phagocytize over 100 bacteria, digest them and extrude residues, and then survive for many more months. Macrophages are also capable of ingesting large particles, including red blood cells and malarial parasites. They increase several-fold in size and transform into macrophages that are characteristic of the tissue they have entered, surviving in tissues for several months.

Mononuclear phagocytes are essential in defending the body against invasion by foreign pathogens, particularly intracellular microorganisms such as M. tuberculosis, listeria, leishmania and toxoplasma. Macrophages can also control the growth of tumorous cells, via both phagocytosis and secretion of hydrolytic enzymes. Another important function of macrophages is that of processing

antigen and presenting them in a biochemically modified form to lymphocytes.

The immune system responds to invading microorganisms in two major ways: antibody production and cell mediated responses. Antibodies are immunoglobulin proteins produced by B-lymphocytes which bind to specific antigens and cause inactivation or promote destruction of the antigen by other cells. Cell-mediated immune responses involve T-lymphocytes (T cells) that react with foreign antigen on the surface of infected host cells. Depending on the type of T cell, the infected cell is either killed or signals are secreted which activate macrophages and other cells to destroy the infected cell (Paul, supra).

T-lymphocytes originate in the bone marrow or liver in fetuses. Precursor cells migrate via the blood to the thymus, where they are processed to mature into T-lymphocytes. This processing is crucial because of positive and negative selection of T cells that will react with foreign antigen and not with self molecules. After processing, T cells continuously circulate in the blood and secondary lymphoid tissues, such as lymph nodes, spleen, certain epithelium-associated tissues in the gastrointestinal tract, respiratory tract and skin. When T-lymphocytes are presented with the complementary antigen, they are stimulated to proliferate and release large numbers of activated T cells into the lymph system and the blood system. These activated T cells can survive and circulate for several days. At the same time, T memory cells are created, which remain in the lymphoid tissue for months or years. Upon subsequent exposure to that specific antigen, these memory cells will respond more rapidly and with a stronger response than induced by the original antigen. This creates an "immunological memory" that can provide immunity for years.

There are two major types of T cells: cytotoxic T cells destroy infected host cells, and helper T cells activate other white blood cells via chemical signals. One class of helper cell, $T_H1$, activates macrophages to destroy ingested microorganisms, while another, $T_H2$, stimulates the production of antibodies by B cells.

Cytotoxic T cells directly attack the infected target cell. In virus-infected cells, peptides derived from viral proteins are generated by the proteasome. These peptides are transported into the ER by the transporter associated with antigen processing (TAP) (Pamer, E. and P. Cresswell (1998) Annu. Rev. Immunol. 16:323-358). Once inside the ER, the peptides bind MHC I chains, and the peptide/MHC I complex is transported to the cell surface. Receptors on the surface of T cells bind to antigen presented on cell surface MHC molecules. Once activated by binding to antigen, T cells secrete γ-interferon, a signal molecule that induces the expression of genes necessary for presenting viral (or other) antigens to cytotoxic T cells. Cytotoxic T cells kill the infected cell by stimulating programmed cell death.

Helper T cells constitute up to 75% of the total T cell population. They regulate the immune functions by producing a variety of lymphokines that act on other cells in the immune system and on

bone marrow. Among these lymphokines are: interleukins-2,3,4,5,6; granulocyte-monocyte colony stimulating factor, and γ-interferon.

Helper T cells are required for most B cells to respond to antigen. When an activated helper cell contacts a B cell, its centrosome and Golgi apparatus become oriented toward the B cell, aiding the directing of signal molecules, such as transmembrane-bound protein called CD40 ligand, onto the B cell surface to interact with the CD40 transmembrane protein. Secreted signals also help B cells to proliferate and mature and, in some cases, to switch the class of antibody being produced.

B-lymphocytes (B cells) produce antibodies which react with specific antigenic proteins presented by pathogens. Once activated, B cells become filled with extensive rough endoplasmic reticulum and are known as plasma cells. As with T cells, interaction of B cells with antigen stimulates proliferation of only those B cells which produce antibody specific to that antigen. There are five classes of antibodies, known as immunoglobulins, which together comprise about 20% of total plasma protein. Each class mediates a characteristic biological response after antigen binding. Upon activation by specific antigen B cells switch from making membrane-bound antibody to secretion of that antibody.

Antibodies, or immunoglobulins (Ig), are the founding members of the Ig superfamily and the central components of the humoral immune response. Antibodies are either expressed on the surface of B cells or secreted by B cells into the circulation. Antibodies bind and neutralize blood-borne foreign antigens. The prototypical antibody is a tetramer consisting of two identical heavy polypeptide chains (H-chains) and two identical light polypeptide chains (L-chains) interlinked by disulfide bonds. This arrangement confers the characteristic Y-shape to antibody molecules. Antibodies are classified based on their H-chain composition. The five antibody classes, IgA, IgD, IgE, IgG and IgM, are defined by the α, δ, ϵ, γ, and μ H-chain types. There are two types of L-chains, κ and λ, either of which may associate as a pair with any H-chain pair. IgG, the most common class of antibody found in the circulation, is tetrameric, while the other classes of antibodies are generally variants or multimers of this basic structure.

H-chains and L-chains each contain an N-terminal variable region and a C-terminal constant region. Both H-chains and L-chains contain repeated Ig domains. For example, a typical H-chain contains four Ig domains, three of which occur within the constant region and one of which occurs within the variable region and contributes to the formation of the antigen recognition site. Likewise, a typical L-chain contains two Ig domains, one of which occurs within the constant region and one of which occurs within the variable region. In addition, H chains such as μ have been shown to associate with other polypeptides during differentiation of the B cell.

Antibodies can be described in terms of their two main functional domains. Antigen recognition is mediated by the Fab (antigen binding fragment) region of the antibody, while effector

functions are mediated by the Fc (crystallizable fragment) region. Binding of antibody to an antigen, such as a bacterium, triggers the destruction of the antigen by phagocytic white blood cells such as macrophages and neutrophils. These cells express surface receptors that specifically bind to the antibody Fc region and allow the phagocytic cells to engulf, ingest, and degrade the antibody-bound antigen. The Fc receptors expressed by phagocytic cells are single-pass transmembrane glycoproteins of about 300 to 400 amino acids (Sears, D.W. et al. (1990) J. Immunol. 144:371-378). The extracellular portion of the Fc receptor typically contains two or three Ig domains.

Diseases which cause over- or under-abundance of any one type of leukocyte usually result in the entire immune defense system becoming involved. A well-known autoimmune disease is AIDS (Acquired Immunodeficiency Syndrome) where the number of helper T cells is depleted, leaving the patient susceptible to infection by microorganisms and parasites. Another widespread medical condition attributable to the immune system is that of allergic reactions to certain antigens. Allergic reactions include: hay fever, asthma, anaphylaxis, and urticaria (hives). Leukemias are an excess production of white blood cells, to the point where a major portion of the body's metabolic resources are directed solely at proliferation of white blood cells, leaving other tissues to starve. Leukopenia or agranulocytosis occurs when the bone marrow stops producing white blood cells. This leaves the body unprotected against foreign microorganisms, including those which normally inhabit skin, mucous membranes, and gastrointestinal tract. If all white blood cell production stops completely, infection will occur within two days and death may follow only 1 to 4 days later.

Impaired phagocytosis occurs in several diseases, including monocytic leukemia, systemic lupus, and granulomatous disease. In such a situation, macrophages can phagocytize normally, but the enveloped organism is not killed. A defect in the plasma membrane enzyme which converts oxygen to lethally reactive forms results in abscess formation in liver, lungs, spleen, lymph nodes, and beneath the skin. Eosinophilia is an excess of eosinophils commonly observed in patients with allergies (hay fever, asthma), allergic reactions to drugs, rheumatoid arthritis, and cancers (Hodgkin's disease, lung, and liver cancer) (Isselbacher, K.J. et al. (1994) Harrison's Principles of Internal Medicine, McGraw-Hill, Inc., New York NY).

Host defense is further augmented by the complement system. The complement system serves as an effector system and is involved in infectious agent recognition. It can function as an independent immune network or in conjunction with other humoral immune responses. The complement system is comprised of numerous plasma and membrane proteins that act in a cascade of reaction sequences whereby one component activates the next. The result is a rapid and amplified response to infection through either an inflammatory response or increased phagocytosis.

The complement system has more than 30 protein components which can be divided into functional groupings including modified serine proteases, membrane-binding proteins and regulators

of complement activation. Activation occurs through two different pathways the classical and the alternative. Both pathways serve to destroy infectious agents through distinct triggering mechanisms that eventually merge with the involvement of the component C3.

The classical pathway requires antibody binding to infectious agent antigens. The antibodies serve to define the target and initiate the complement system cascade, culminating in the destruction of the infectious agent. In this pathway, since the antibody guides initiation of the process, the complement can be seen as an effector arm of the humoral immune system.

The alternative pathway of the complement system does not require the presence of pre-existing antibodies for targeting infectious agent destruction. Rather, this pathway, through low levels of an activated component, remains constantly primed and provides surveillance in the non-immune host to enable targeting and destruction of infectious agents. In this case foreign material triggers the cascade, thereby facilitating phagocytosis or lysis (Paul, supra, pp.918-919).

Another important component of host defense is the process of inflammation. Inflammatory responses are divided into four categories on the basis of pathology and include allergic inflammation, cytotoxic antibody mediated inflammation, immune complex mediated inflammation and monocyte mediated inflammation. Inflammation manifests as a combination of each of these forms with one predominating.

Allergic acute inflammation is observed in individuals wherein specific antigens stimulate IgE antibody production. Mast cells and basophils are subsequently activated by the attachment of antigen-IgE complexes, resulting in the release of cytoplasmic granule contents such as histamine. The products of activated mast cells can increase vascular permeability and constrict the smooth muscle of breathing passages, resulting in anaphylaxis or asthma. Acute inflammation is also mediated by cytotoxic antibodies and can result in the destruction of tissue through the binding of complement-fixing antibodies to cells. The responsible antibodies are of the IgG or IgM types. Resultant clinical disorders include autoimmune hemolytic anemia and thrombocytopenia as associated with systemic lupus erythematosis.

Immune complex mediated acute inflammation involves the IgG or IgM antibody types which combine with antigen to activate the complement cascade. When such immune complexes bind to neutrophils and macrophages they activate the respiratory burst to form protein- and vessel-damaging agents such as hydrogen peroxide, hydroxyl radical, hypochlorous acid, and chloramines. Clinical manifestations include rheumatoid arthritis and systemic lupus erythematosus.

In chronic inflammation or delayed-type hypersensitivity, macrophages are activated and process antigen for presentation to T cells that subsequently produce lymphokines and monokines. This type of inflammatory response is likely important for defense against intracellular parasites and certain viruses. Clinical associations include, granulomatous disease, tuberculosis, leprosy, and

sarcoidosis (Paul, W.E., supra, pp.1017-1018).

**Extracellular Information Transmission Molecules**

Intercellular communication is essential for the growth and survival of multicellular organisms, and in particular, for the function of the endocrine, nervous, and immune systems. In addition, intercellular communication is critical for developmental processes such as tissue construction and organogenesis, in which cell proliferation, cell differentiation, and morphogenesis must be spatially and temporally regulated in a precise and coordinated manner. Cells communicate with one another through the secretion and uptake of diverse types of signaling molecules such as hormones, growth factors, neuropeptides, and cytokines.

Hormones

Hormones are signaling molecules that coordinately regulate basic physiological processes from embryogenesis throughout adulthood. These processes include metabolism, respiration, reproduction, excretion, fetal tissue differentiation and organogenesis, growth and development, homeostasis, and the stress response. Hormonal secretions and the nervous system are tightly integrated and interdependent. Hormones are secreted by endocrine glands, primarily the hypothalamus and pituitary, the thyroid and parathyroid, the pancreas, the adrenal glands, and the ovaries and testes.

The secretion of hormones into the circulation is tightly controlled. Hormones are often secreted in diurnal, pulsatile, and cyclic patterns. Hormone secretion is regulated by perturbations in blood biochemistry, by other upstream-acting hormones, by neural impulses, and by negative feedback loops. Blood hormone concentrations are constantly monitored and adjusted to maintain optimal, steady-state levels. Once secreted, hormones act only on those target cells that express specific receptors.

Most disorders of the endocrine system are caused by either hyposecretion or hypersecretion of hormones. Hyposecretion often occurs when a hormone's gland of origin is damaged or otherwise impaired. Hypersecretion often results from the proliferation of tumors derived from hormone-secreting cells. Inappropriate hormone levels may also be caused by defects in regulatory feedback loops or in the processing of hormone precursors. Endocrine malfunction may also occur when the target cell fails to respond to the hormone.

Hormones can be classified biochemically as polypeptides, steroids, eicosanoids, or amines. Polypeptides, which include diverse hormones such as insulin and growth hormone, vary in size and function and are often synthesized as inactive precursors that are processed intracellularly into mature, active forms. Amines, which include epinephrine and dopamine, are amino acid derivatives that function in neuroendocrine signaling. Steroids, which include the cholesterol-derived hormones estrogen and testosterone, function in sexual development and reproduction. Eicosanoids, which

include prostaglandins and prostacyclins, are fatty acid derivatives that function in a variety of processes. Most polypeptides and some amines are soluble in the circulation where they are highly susceptible to proteolytic degradation within seconds after their secretion. Steroids and lipids are insoluble and must be transported in the circulation by carrier proteins. The following discussion will focus primarily on polypeptide hormones.

Hormones secreted by the hypothalamus and pituitary gland play a critical role in endocrine function by coordinately regulating hormonal secretions from other endocrine glands in response to neural signals. Hypothalamic hormones include thyrotropin-releasing hormone, gonadotropin-releasing hormone, somatostatin, growth-hormone releasing factor, corticotropin-releasing hormone, substance P, dopamine, and prolactin-releasing hormone. These hormones directly regulate the secretion of hormones from the anterior lobe of the pituitary. Hormones secreted by the anterior pituitary include adrenocorticotropic hormone (ACTH), melanocyte-stimulating hormone, somatotropic hormones such as growth hormone and prolactin, glycoprotein hormones such as thyroid-stimulating hormone, luteinizing hormone (LH), and follicle-stimulating hormone (FSH), β-lipotropin, and β-endorphins. These hormones regulate hormonal secretions from the thyroid, pancreas, and adrenal glands, and act directly on the reproductive organs to stimulate ovulation and spermatogenesis. The posterior pituitary synthesizes and secretes antidiuretic hormone (ADH, vasopressin) and oxytocin.

Disorders of the hypothalamus and pituitary often result from lesions such as primary brain tumors, adenomas, infarction associated with pregnancy, hypophysectomy, aneurysms, vascular malformations, thrombosis, infections, immunological disorders, and complications due to head trauma. Such disorders have profound effects on the function of other endocrine glands. Disorders associated with hypopituitarism include hypogonadism, Sheehan syndrome, diabetes insipidus, Kallman's disease, Hand-Schuller-Christian disease, Letterer-Siwe disease, sarcoidosis, empty sella syndrome, and dwarfism. Disorders associated with hyperpituitarism include acromegaly, giantism, and syndrome of inappropriate ADH secretion (SIADH), often caused by benign adenomas.

Hormones secreted by the thyroid and parathyroid primarily control metabolic rates and the regulation of serum calcium levels, respectively. Thyroid hormones include calcitonin, somatostatin, and thyroid hormone. The parathyroid secretes parathyroid hormone. Disorders associated with hypothyroidism include goiter, myxedema, acute thyroiditis associated with bacterial infection, subacute thyroiditis associated with viral infection, autoimmune thyroiditis (Hashimoto's disease), and cretinism. Disorders associated with hyperthyroidism include thyrotoxicosis and its various forms, Grave's disease, pretibial myxedema, toxic multinodular goiter, thyroid carcinoma, and Plummer's disease. Disorders associated with hyperparathyroidism include Conn disease (chronic hypercalemia) leading to bone resorption and parathyroid hyperplasia.

Hormones secreted by the pancreas regulate blood glucose levels by modulating the rates of

carbohydrate, fat, and protein metabolism. Pancreatic hormones include insulin, glucagon, amylin, γ-aminobutyric acid, gastrin, somatostatin, and pancreatic polypeptide. The principal disorder associated with pancreatic dysfunction is diabetes mellitus caused by insufficient insulin activity. Diabetes mellitus is generally classified as either Type I (insulin-dependent, juvenile diabetes) or Type II (non-insulin-dependent, adult diabetes). The treatment of both forms by insulin replacement therapy is well known. Diabetes mellitus often leads to acute complications such as hypoglycemia (insulin shock), coma, diabetic ketoacidosis, lactic acidosis, and chronic complications leading to disorders of the eye, kidney, skin, bone, joint, cardiovascular system, nervous system, and to decreased resistance to infection.

The anatomy, physiology, and diseases related to hormonal function are reviewed in McCance, K.L. and S.E. Huether (1994) Pathophysiology: The Biological Basis for Disease in Adults and Children, Mosby-Year Book, Inc., St. Louis MO; Greenspan, F.S. and J.D. Baxter (1994) Basic and Clinical Endocrinology, Appleton and Lange, East Norwalk CT.

Growth Factors

Growth factors are secreted proteins that mediate intercellular communication. Unlike hormones, which travel great distances via the circulatory system, most growth factors are primarily local mediators that act on neighboring cells. Most growth factors contain a hydrophobic N-terminal signal peptide sequence which directs the growth factor into the secretory pathway. Most growth factors also undergo post-translational modifications within the secretory pathway. These modifications can include proteolysis, glycosylation, phosphorylation, and intramolecular disulfide bond formation. Once secreted, growth factors bind to specific receptors on the surfaces of neighboring target cells, and the bound receptors trigger intracellular signal transduction pathways. These signal transduction pathways elicit specific cellular responses in the target cells. These responses can include the modulation of gene expression and the stimulation or inhibition of cell division, cell differentiation, and cell motility.

Growth factors fall into at least two broad and overlapping classes. The broadest class includes the large polypeptide growth factors, which are wide-ranging in their effects. These factors include epidermal growth factor (EGF), fibroblast growth factor (FGF), transforming growth factor-β (TGF-β), insulin-like growth factor (IGF), nerve growth factor (NGF), and platelet-derived growth factor (PDGF), each defining a family of numerous related factors. The large polypeptide growth factors, with the exception of NGF, act as mitogens on diverse cell types to stimulate wound healing, bone synthesis and remodeling, extracellular matrix synthesis, and proliferation of epithelial, epidermal, and connective tissues. Members of the TGF-β, EGF, and FGF families also function as inductive signals in the differentiation of embryonic tissue. NGF functions specifically as a neurotrophic factor, promoting neuronal growth and differentiation.

Another class of growth factors includes the hematopoietic growth factors, which are narrow in their target specificity. These factors stimulate the proliferation and differentiation of blood cells such as B-lymphocytes, T-lymphocytes, erythrocytes, platelets, eosinophils, basophils, neutrophils, macrophages, and their stem cell precursors. These factors include the colony-stimulating factors (G-CSF, M-CSF, GM-CSF, and CSF1-3), erythropoietin, and the cytokines. The cytokines are specialized hematopoietic factors secreted by cells of the immune system and are discussed in detail below.

Growth factors play critical roles in neoplastic transformation of cells in vitro and in tumor progression in vivo. Overexpression of the large polypeptide growth factors promotes the proliferation and transformation of cells in culture. Inappropriate expression of these growth factors by tumor cells in vivo may contribute to tumor vascularization and metastasis. Inappropriate activity of hematopoietic growth factors can result in anemias, leukemias, and lymphomas. Moreover, growth factors are both structurally and functionally related to oncoproteins, the potentially cancer-causing products of proto-oncogenes. Certain FGF and PDGF family members are themselves homologous to oncoproteins, whereas receptors for some members of the EGF, NGF, and FGF families are encoded by proto-oncogenes. Growth factors also affect the transcriptional regulation of both proto-oncogenes and oncosuppressor genes (Pimentel, E. (1994) Handbook of Growth Factors, CRC Press, Ann Arbor MI; McKay, I. and I. Leigh, eds. (1993) Growth Factors: A Practical Approach, Oxford University Press, New York NY; Habenicht, A., ed. (1990) Growth Factors, Differentiation Factors, and Cytokines, Springer-Verlag, New York NY).

In addition, some of the large polypeptide growth factors play crucial roles in the induction of the primordial germ layers in the developing embryo. This induction ultimately results in the formation of the embryonic mesoderm, ectoderm, and endoderm which in turn provide the framework for the entire adult body plan. Disruption of this inductive process would be catastrophic to embryonic development.

Small Peptide Factors - Neuropeptides and Vasomediators

Neuropeptides and vasomediators (NP/VM) comprise a family of small peptide factors, typically of 20 amino acids or less. These factors generally function in neuronal excitation and inhibition of vasoconstriction/vasodilation, muscle contraction, and hormonal secretions from the brain and other endocrine tissues. Included in this family are neuropeptides and neuropeptide hormones such as bombesin, neuropeptide Y, neurotensin, neuromedin N, melanocortins, opioids, galanin, somatostatin, tachykinins, urotensin II and related peptides involved in smooth muscle stimulation, vasopressin, vasoactive intestinal peptide, and circulatory system-borne signaling molecules such as angiotensin, complement, calcitonin, endothelins, formyl-methionyl peptides, glucagon, cholecystokinin, gastrin, and many of the peptide hormones discussed above. NP/VMs can transduce signals directly, modulate the activity or release of other neurotransmitters and hormones, and

act as catalytic enzymes in signaling cascades. The effects of NP/VMs range from extremely brief to long-lasting. (Reviewed in Martin, C.R. et al. (1985) Endocrine Physiology, Oxford University Press, New York NY, pp. 57-62.)

Cytokines

Cytokines comprise a family of signaling molecules that modulate the immune system and the inflammatory response. Cytokines are usually secreted by leukocytes, or white blood cells, in response to injury or infection. Cytokines function as growth and differentiation factors that act primarily on cells of the immune system such as B- and T-lymphocytes, monocytes, macrophages, and granulocytes. Like other signaling molecules, cytokines bind to specific plasma membrane receptors and trigger intracellular signal transduction pathways which alter gene expression patterns. There is considerable potential for the use of cytokines in the treatment of inflammation and immune system disorders.

Cytokine structure and function have been extensively characterized in vitro. Most cytokines are small polypeptides of about 30 kilodaltons or less. Over 50 cytokines have been identified from human and rodent sources. Examples of cytokine subfamilies include the interferons (IFN-$\alpha$, -$\beta$, and -$\gamma$), the interleukins (IL1-IL13), the tumor necrosis factors (TNF-$\alpha$ and -$\beta$), and the chemokines. Many cytokines have been produced using recombinant DNA techniques, and the activities of individual cytokines have been determined in vitro. These activities include regulation of leukocyte proliferation, differentiation, and motility.

The activity of an individual cytokine in vitro may not reflect the full scope of that cytokine's activity in vivo. Cytokines are not expressed individually in vivo but are instead expressed in combination with a multitude of other cytokines when the organism is challenged with a stimulus. Together, these cytokines collectively modulate the immune response in a manner appropriate for that particular stimulus. Therefore, the physiological activity of a cytokine is determined by the stimulus itself and by complex interactive networks among co-expressed cytokines which may demonstrate both synergistic and antagonistic relationships.

Chemokines comprise a cytokine subfamily with over 30 members. (Reviewed in Wells, T. N.C. and M.C. Peitsch (1997) J. Leukoc. Biol. 61:545-550.) Chemokines were initially identified as chemotactic proteins that recruit monocytes and macrophages to sites of inflammation. Recent evidence indicates that chemokines may also play key roles in hematopoiesis and HIV-1 infection. Chemokines are small proteins which range from about 6-15 kilodaltons in molecular weight. Chemokines are further classified as C, CC, CXC, or $CX_3C$ based on the number and position of critical cysteine residues. The CC chemokines, for example, each contain a conserved motif consisting of two consecutive cysteines followed by two additional cysteines which occur downstream at 24- and 16-residue intervals, respectively (ExPASy PROSITE database, documents PS00472 and PDOC00434). The presence and spacing of these four cysteine residues are highly conserved, whereas the intervening

residues diverge significantly. However, a conserved tyrosine located about 15 residues downstream of the cysteine doublet seems to be important for chemotactic activity. Most of the human genes encoding CC chemokines are clustered on chromosome 17, although there are a few examples of CC chemokine genes that map elsewhere. Other chemokines include lymphotactin (C chemokine); macrophage chemotactic and activating factor (MCAF/MCP-1; CC chemokine); platelet factor 4 and IL-8 (CXC chemokines); and fractalkine and neurotractin ($CX_3C$ chemokines). (Reviewed in Luster, A.D. (1998) N. Engl. J. Med. 338:436-445.)

**Receptor Molecules**

SEQ ID NO:6 and SEQ ID NO:7 encode, for example, receptor molecules.

The term receptor describes proteins that specifically recognize other molecules. The category is broad and includes proteins with a variety of functions. The bulk of receptors are cell surface proteins which bind extracellular ligands and produce cellular responses in the areas of growth, differentiation, endocytosis, and immune response. Other receptors facilitate the selective transport of proteins out of the endoplasmic reticulum and localize enzymes to particular locations in the cell. The term may also be applied to proteins which act as receptors for ligands with known or unknown chemical composition and which interact with other cellular components. For example, the steroid hormone receptors bind to and regulate transcription of DNA.

Regulation of cell proliferation, differentiation, and migration is important for the formation and function of tissues. Regulatory proteins such as growth factors coordinately control these cellular processes and act as mediators in cell-cell signaling pathways. Growth factors are secreted proteins that bind to specific cell-surface receptors on target cells. The bound receptors trigger intracellular signal transduction pathways which activate various downstream effectors that regulate gene expression, cell division, cell differentiation, cell motility, and other cellular processes.

Cell surface receptors are typically integral plasma membrane proteins. These receptors recognize hormones such as catecholamines; peptide hormones; growth and differentiation factors; small peptide factors such as thyrotropin-releasing hormone; galanin, somatostatin, and tachykinins; and circulatory system-borne signaling molecules. Cell surface receptors on immune system cells recognize antigens, antibodies, and major histocompatibility complex (MHC)-bound peptides. Other cell surface receptors bind ligands to be internalized by the cell. This receptor-mediated endocytosis functions in the uptake of low density lipoproteins (LDL), transferrin, glucose- or mannose-terminal glycoproteins, galactose-terminal glycoproteins, immunoglobulins, phosphovitellogenins, fibrin, proteinase-inhibitor complexes, plasminogen activators, and thrombospondin (Lodish, H. et al. (1995) Molecular Cell Biology, Scientific American Books, New York NY, p. 723; Mikhailenko, I. et al. (1997) J. Biol. Chem. 272:6784-6791).

Receptor Protein Kinases

Many growth factor receptors, including receptors for epidermal growth factor, platelet-derived growth factor, fibroblast growth factor, as well as the growth modulator α-thrombin, contain intrinsic protein kinase activities. When growth factor binds to the receptor, it triggers the autophosphorylation of a serine, threonine, or tyrosine residue on the receptor. These phosphorylated sites are recognition sites for the binding of other cytoplasmic signaling proteins. These proteins participate in signaling pathways that eventually link the initial receptor activation at the cell surface to the activation of a specific intracellular target molecule. In the case of tyrosine residue autophosphorylation, these signaling proteins contain a common domain referred to as a Src homology (SH) domain. SH2 domains and SH3 domains are found in phospholipase C-γ, PI-3-K p85 regulatory subunit, Ras-GTPase activating protein, and $pp60^{c\text{-}src}$ (Lowenstein, E.J. et al. (1992) Cell 70:431-442). The cytokine family of receptors share a different common binding domain and include transmembrane receptors for growth hormone (GH), interleukins, erythropoietin, and prolactin.

Other receptors and second messenger-binding proteins have intrinsic serine/threonine protein kinase activity. These include activin/TGF-β/BMP-superfamily receptors, calcium- and diacylglycerol-activated/phospholipid-dependant protein kinase (PK-C), and RNA-dependant protein kinase (PK-R). In addition, other serine/threonine protein kinases, including nematode Twitchin, have fibronectin-like, immunoglobulin C2-like domains.

G-Protein Coupled Receptors

G-protein coupled receptors (GPCRs) are integral membrane proteins characterized by the presence of seven hydrophobic transmembrane domains which span the plasma membrane and form a bundle of antiparallel alpha (α) helices. These proteins range in size from under 400 to over 1000 amino acids (Strosberg, A.D. (1991) Eur. J. Biochem. 196:1-10; Coughlin, S.R. (1994) Curr. Opin. Cell Biol. 6:191-197). The amino-terminus of the GPCR is extracellular, of variable length and often glycosylated; the carboxy-terminus is cytoplasmic and generally phosphorylated. Extracellular loops of the GPCR alternate with intracellular loops and link the transmembrane domains. The most conserved domains of GPCRs are the transmembrane domains and the first two cytoplasmic loops. The transmembrane domains account for structural and functional features of the receptor. In most cases, the bundle of α helices forms a binding pocket. In addition, the extracellular N-terminal segment or one or more of the three extracellular loops may also participate in ligand binding. Ligand binding activates the receptor by inducing a conformational change in intracellular portions of the receptor. The activated receptor, in turn, interacts with an intracellular heterotrimeric guanine nucleotide binding (G) protein complex which mediates further intracellular signaling activities, generally the production of second messengers such as cyclic AMP (cAMP), phospholipase C, inositol triphosphate, or interactions with ion channel proteins (Baldwin, J.M. (1994) Curr. Opin. Cell Biol. 6:180-190).

GPCRs include those for acetylcholine, adenosine, epinephrine and norepinephrine, bombesin, bradykinin, chemokines, dopamine, endothelin, γ-aminobutyric acid (GABA), follicle-stimulating hormone (FSH), glutamate, gonadotropin-releasing hormone (GnRH), hepatocyte growth factor, histamine, leukotrienes, melanocortins, neuropeptide Y, opioid peptides, opsins, prostanoids, serotonin, somatostatin, tachykinins, thrombin, thyrotropin-releasing hormone (TRH), vasoactive intestinal polypeptide family, vasopressin and oxytocin, and orphan receptors.

GPCR mutations, which may cause loss of function or constitutive activation, have been associated with numerous human diseases (Coughlin, supra). For instance, retinitis pigmentosa may arise from mutations in the rhodopsin gene. Rhodopsin is the retinal photoreceptor which is located within the discs of the eye rod cell. Parma, J. et al. (1993, Nature 365:649-651) report that somatic activating mutations in the thyrotropin receptor cause hyperfunctioning thyroid adenomas and suggest that certain GPCRs susceptible to constitutive activation may behave as protooncogenes.

Nuclear Receptors

Nuclear receptors bind small molecules such as hormones or second messengers, leading to increased receptor-binding affinity to specific chromosomal DNA elements. In addition the affinity for other nuclear proteins may also be altered. Such binding and protein-protein interactions may regulate and modulate gene expression. Examples of such receptors include the steroid hormone receptors family, the retinoic acid receptors family, and the thyroid hormone receptors family.

Ligand-Gated Receptor Ion Channels

Ligand-gated receptor ion channels fall into two categories. The first category, extracellular ligand-gated receptor ion channels (ELGs), rapidly transduce neurotransmitter-binding events into electrical signals, such as fast synaptic neurotransmission. ELG function is regulated by post-translational modification. The second category, intracellular ligand-gated receptor ion channels (ILGs), are activated by many intracellular second messengers and do not require post-translational modification(s) to effect a channel-opening response.

ELGs depolarize excitable cells to the threshold of action potential generation. In non-excitable cells, ELGs permit a limited calcium ion-influx during the presence of agonist. ELGs include channels directly gated by neurotransmitters such as acetylcholine, L-glutamate, glycine, ATP, serotonin, GABA, and histamine. ELG genes encode proteins having strong structural and functional similarities. ILGs are encoded by distinct and unrelated gene families and include receptors for cAMP, cGMP, calcium ions, ATP, and metabolites of arachidonic acid.

Macrophage Scavenger Receptors

Macrophage scavenger receptors with broad ligand specificity may participate in the binding of low density lipoproteins (LDL) and foreign antigens. Scavenger receptors types I and II are trimeric membrane proteins with each subunit containing a small N-terminal intracellular domain, a

transmembrane domain, a large extracellular domain, and a C-terminal cysteine-rich domain. The extracellular domain contains a short spacer domain, an α-helical coiled-coil domain, and a triple helical collagenous domain. These receptors have been shown to bind a spectrum of ligands, including chemically modified lipoproteins and albumin, polyribonucleotides, polysaccharides, phospholipids, and asbestos (Matsumoto, A. et al. (1990) Proc. Natl. Acad. Sci. USA 87:9133-9137; Elomaa, O. et al. (1995) Cell 80:603-609). The scavenger receptors are thought to play a key role in atherogenesis by mediating uptake of modified LDL in arterial walls, and in host defense by binding bacterial endotoxins, bacteria, and protozoa.

T-Cell Receptors

T cells play a dual role in the immune system as effectors and regulators, coupling antigen recognition with the transmission of signals that induce cell death in infected cells and stimulate proliferation of other immune cells. Although a population of T cells can recognize a wide range of different antigens, an individual T cell can only recognize a single antigen and only when it is presented to the T cell receptor (TCR) as a peptide complexed with a major histocompatibility molecule (MHC) on the surface of an antigen presenting cell. The TCR on most T cells consists of immunoglobulin-like integral membrane glycoproteins containing two polypeptide subunits, α and β, of similar molecular weight. Both TCR subunits have an extracellular domain containing both variable and constant regions, a transmembrane domain that traverses the membrane once, and a short intracellular domain (Saito, H. et al. (1984) Nature 309:757-762). The genes for the TCR subunits are constructed through somatic rearrangement of different gene segments. Interaction of antigen in the proper MHC context with the TCR initiates signaling cascades that induce the proliferation, maturation, and function of cellular components of the immune system (Weiss, A. (1991) Annu. Rev. Genet. 25:487-510). Rearrangements in TCR genes and alterations in TCR expression have been noted in lymphomas, leukemias, autoimmune disorders, and immunodeficiency disorders (Aisenberg, A.C. et al. (1985) N. Engl. J. Med. 313:529-533; Weiss, supra).

**Intracellular Signaling Molecules**

SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, and SEQ ID NO:12 encode, for example, intracellular signaling molecules.

Intracellular signaling is the general process by which cells respond to extracellular signals (hormones, neurotransmitters, growth and differentiation factors, etc.) through a cascade of biochemical reactions that begins with the binding of a signaling molecule to a cell membrane receptor and ends with the activation of an intracellular target molecule. Intermediate steps in the process involve the activation of various cytoplasmic proteins by phosphorylation via protein kinases, and their deactivation by protein phosphatases, and the eventual translocation of some of these

activated proteins to the cell nucleus where the transcription of specific genes is triggered. The intracellular signaling process regulates all types of cell functions including cell proliferation, cell differentiation, and gene transcription, and involves a diversity of molecules including protein kinases and phosphatases, and second messenger molecules, such as cyclic nucleotides, calcium-calmodulin, inositol, and various mitogens, that regulate protein phosphorylation.

Protein Phosphorylation

Protein kinases and phosphatases play a key role in the intracellular signaling process by controlling the phosphorylation and activation of various signaling proteins. The high energy phosphate for this reaction is generally transferred from the adenosine triphosphate molecule (ATP) to a particular protein by a protein kinase and removed from that protein by a protein phosphatase. Protein kinases are roughly divided into two groups: those that phosphorylate tyrosine residues (protein tyrosine kinases, PTK) and those that phosphorylate serine or threonine residues (serine/threonine kinases, STK). A few protein kinases have dual specificity for serine/threonine and tyrosine residues. Almost all kinases contain a conserved 250-300 amino acid catalytic domain containing specific residues and sequence motifs characteristic of the kinase family (Hardic, G. and S. Hanks (1995) The Protein Kinase Facts Books, Vol I:7-20, Academic Press, San Diego CA).

STKs include the second messenger dependent protein kinases such as the cyclic-AMP dependent protein kinases (PKA), involved in mediating hormone-induced cellular responses; calcium-calmodulin (CaM) dependent protein kinases, involved in regulation of smooth muscle contraction, glycogen breakdown, and neurotransmission; and the mitogen-activated protein kinases (MAP) which mediate signal transduction from the cell surface to the nucleus via phosphorylation cascades. Altered PKA expression is implicated in a variety of disorders and diseases including cancer, thyroid disorders, diabetes, atherosclerosis, and cardiovascular disease (Isselbacher, K.J. et al. (1994) Harrison's Principles of Internal Medicine, McGraw-Hill, New York NY, pp. 416-431, 1887).

PTKs are divided into transmembrane, receptor PTKs and nontransmembrane, non-receptor PTKs. Transmembrane PTKs are receptors for most growth factors. Non-receptor PTKs lack transmembrane regions and, instead, form complexes with the intracellular regions of cell surface receptors. Receptors that function through non-receptor PTKs include those for cytokines and hormones (growth hormone and prolactin) and antigen-specific receptors on T and B lymphocytes. Many of these PTKs were first identified as the products of mutant oncogenes in cancer cells in which their activation was no longer subject to normal cellular controls. In fact, about one third of the known oncogenes encode PTKs, and it is well known that cellular transformation (oncogenesis) is often accompanied by increased tyrosine phosphorylation activity (Charbonneau, H. and N.K. Tonks (1992) Annu. Rev. Cell Biol. 8:463-493).

An additional family of protein kinases previously thought to exist only in procaryotes is the

histidine protein kinase family (HPK). HPKs bear little homology with mammalian STKs or PTKs but have distinctive sequence motifs of their own (Davie, J.R. et al. (1995) J. Biol. Chem. 270:19861-19867). A histidine residue in the N-terminal half of the molecule (region I) is an autophosphorylation site. Three additional motifs located in the C-terminal half of the molecule include an invariant asparagine residue in region II and two glycine-rich loops characteristic of nucleotide binding domains in regions III and IV. Recently a branched chain alpha-ketoacid dehydrogenase kinase has been found with characteristics of HPK in rat (Davie, supra).

Protein phosphatases regulate the effects of protein kinases by removing phosphate groups from molecules previously activated by kinases. The two principal categories of protein phosphatases are the protein (serine/threonine) phosphatases (PPs) and the protein tyrosine phosphatases (PTPs). PPs dephosphorylate phosphoserine/threonine residues and are important regulators of many cAMP-mediated hormone responses (Cohen, P. (1989) Annu. Rev. Biochem. 58:453-508). PTPs reverse the effects of protein tyrosine kinases and play a significant role in cell cycle and cell signaling processes (Charbonneau, supra). As previously noted, many PTKs are encoded by oncogenes, and oncogenesis is often accompanied by increased tyrosine phosphorylation activity. It is therefore possible that PTPs may prevent or reverse cell transformation and the growth of various cancers by controlling the levels of tyrosine phosphorylation in cells. This hypothesis is supported by studies showing that overexpression of PTPs can suppress transformation in cells, and that specific inhibition of PTPs can enhance cell transformation (Charbonneau, supra).

Phospholipid and Inositol-Phosphate Signaling

Inositol phospholipids (phosphoinositides) are involved in an intracellular signaling pathway that begins with binding of a signaling molecule to a G-protein linked receptor in the plasma membrane. This leads to the phosphorylation of phosphatidylinositol (PI) residues on the inner side of the plasma membrane to the biphosphate state ($PIP_2$) by inositol kinases. Simultaneously, the G-protein linked receptor binding stimulates a trimeric G-protein which in turn activates a phosphoinositide-specific phospholipase C-β. Phospholipase C-β then cleaves $PIP_2$ into two products, inositol triphosphate ($IP_3$) and diacylglycerol. These two products act as mediators for separate signaling events. $IP_3$ diffuses through the plasma membrane to induce calcium release from the endoplasmic reticulum (ER), while diacylglycerol remains in the membrane and helps activate protein kinase C, an STK that phosphorylates selected proteins in the target cell. The calcium response initiated by $IP_3$ is terminated by the dephosphorylation of $IP_3$ by specific inositol phosphatases. Cellular responses that are mediated by this pathway are glycogen breakdown in the liver in response to vasopressin, smooth muscle contraction in response to acetylcholine, and thrombin-induced platelet aggregation.

Cyclic Nucleotide Signaling

Cyclic nucleotides (cAMP and cGMP) function as intracellular second messengers to transduce a variety of extracellular signals including hormones, light, and neurotransmitters. In particular, cyclic-AMP dependent protein kinases (PKA) are thought to account for all of the effects of cAMP in most mammalian cells, including various hormone-induced cellular responses. Visual excitation and the phototransmission of light signals in the eye is controlled by cyclic-GMP regulated, $Ca^{2+}$-specific channels. Because of the importance of cellular levels of cyclic nucleotides in mediating these various responses, regulating the synthesis and breakdown of cyclic nucleotides is an important matter. Thus adenylyl cyclase, which synthesizes cAMP from AMP, is activated to increase cAMP levels in muscle by binding of adrenaline to β-andrenergic receptors, while activation of guanylate cyclase and increased cGMP levels in photoreceptors leads to reopening of the $Ca^{2+}$-specific channels and recovery of the dark state in the eye. In contrast, hydrolysis of cyclic nucleotides by cAMP and cGMP-specific phosphodiesterases (PDEs) produces the opposite of these and other effects mediated by increased cyclic nucleotide levels. PDEs appear to be particularly important in the regulation of cyclic nucleotides, considering the diversity found in this family of proteins. At least seven families of mammalian PDEs (PDE1-7) have been identified based on substrate specificity and affinity, sensitivity to cofactors, and sensitivity to inhibitory drugs (Beavo, J.A. (1995) Physiological Reviews 75:725-48). PDE inhibitors have been found to be particularly useful in treating various clinical disorders. Rolipram, a specific inhibitor of PDE4, has been used in the treatment of depression, and similar inhibitors are undergoing evaluation as anti-inflammatory agents. Theophylline is a nonspecific PDE inhibitor used in the treatment of bronchial asthma and other respiratory diseases (Banner, K.H. and C.P. Page (1995) Eur. Respir. J. 8:996-1000).

G-Protein Signaling

Guanine nucleotide binding proteins (G-proteins) are critical mediators of signal transduction between a particular class of extracellular receptors, the G-protein coupled receptors (GPCR), and intracellular second messengers such as cAMP and $Ca^{2+}$. G-proteins are linked to the cytosolic side of a GPCR such that activation of the GPCR by ligand binding stimulates binding of the G-protein to GTP, inducing an "active" state in the G-protein. In the active state, the G-protein acts as a signal to trigger other events in the cell such as the increase of cAMP levels or the release of $Ca^{2+}$ into the cytosol from the ER, which, in turn, regulate phosphorylation and activation of other intracellular proteins. Recycling of the G-protein to the inactive state involves hydrolysis of the bound GTP to GDP by a GTPase activity in the G-protein. (See Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing, Inc., New York NY, pp.734-759.) Two structurally distinct classes of G-proteins are recognized: heterotrimeric G-proteins, consisting of three different subunits, and monomeric, low molecular weight (LMW), G-proteins consisting of a single polypeptide chain.

The three polypeptide subunits of heterotrimeric G-proteins are the α, β, and γ subunits. The

α subunit binds and hydrolyzes GTP. The β and γ subunits form a tight complex that anchors the protein to the inner side of the plasma membrane. The β subunits, also known as G-β proteins or β transducins, contain seven tandem repeats of the WD-repeat sequence motif, a motif found in many proteins with regulatory functions. Mutations and variant expression of β transducin proteins are linked with various disorders (Neer, E.J. et al. (1994) Nature 371:297-300; Margottin, F. et al. (1998) Mol. Cell 1:565-574).

LMW GTP-proteins are GTPases which regulate cell growth, cell cycle control, protein secretion, and intracellular vesicle interaction. They consist of single polypeptides which, like the α subunit of the heterotrimeric G-proteins, are able to bind and hydrolyze GTP, thus cycling between an inactive and an active state. At least sixty members of the LMW G-protein superfamily have been identified and are currently grouped into the six subfamilies of ras, rho, arf, sar1, ran, and rab. Activated ras genes were initially found in human cancers, and subsequent studies confirmed that ras function is critical in determining whether cells continue to grow or become differentiated. Other members of the LMW G-protein superfamily have roles in signal transduction that vary with the function of the activated genes and the locations of the G-proteins.

Guanine nucleotide exchange factors regulate the activities of LMW G-proteins by determining whether GTP or GDP is bound. GTPase-activating protein (GAP) binds to GTP-ras and induces it to hydrolyze GTP to GDP. In contrast, guanine nucleotide releasing protein (GNRP) binds to GDP-ras and induces the release of GDP and the binding of GTP.

Other regulators of G-protein signaling (RGS) also exist that act primarily by negatively regulating the G-protein pathway by an unknown mechanism (Druey, K.M. et al. (1996) Nature 379:742-746). Some 15 members of the RGS family have been identified. RGS family members are related structurally through similarities in an approximately 120 amino acid region termed the RGS domain and functionally by their ability to inhibit the interleukin (cytokine) induction of MAP kinase in cultured mammalian 293T cells (Druey, supra).

Calcium Signaling Molecules

$Ca^{+2}$ is another second messenger molecule that is even more widely used as an intracellular mediator than cAMP. Two pathways exist by which $Ca^{+2}$ can enter the cytosol in response to extracellular signals: One pathway acts primarily in nerve signal transduction where $Ca^{+2}$ enters a nerve terminal through a voltage-gated $Ca^{+2}$ channel. The second is a more ubiquitous pathway in which $Ca^{+2}$ is released from the ER into the cytosol in response to binding of an extracellular signaling molecule to a receptor. $Ca^{2+}$ directly activates regulatory enzymes, such as protein kinase C, which trigger signal transduction pathways. $Ca^{2+}$ also binds to specific $Ca^{2+}$-binding proteins (CBPs) such as calmodulin (CaM) which then activate multiple target proteins in the cell including enzymes, membrane transport pumps, and ion channels. CaM interactions are involved in a multitude of cellular

processes including, but not limited to, gene regulation, DNA synthesis, cell cycle progression, mitosis, cytokinesis, cytoskeletal organization, muscle contraction, signal transduction, ion homeostasis, exocytosis, and metabolic regulation (Celio, M.R. et al. (1996) Guidebook to Calcium-binding Proteins, Oxford University Press, Oxford, UK, pp. 15-20). Some CBPs can serve as a storage depot for $Ca^{2+}$ in an inactive state. Calsequestrin is one such CBP that is expressed in isoforms specific to cardiac muscle and skeletal muscle. It is suggested that calsequestrin binds $Ca^{2+}$ in a rapidly exchangeable state that is released during $Ca^{2+}$ -signaling conditions (Celio, M.R. et al. (1996) Guidebook to Calcium-binding Proteins, Oxford University Press, New York NY, pp. 222-224).

Cyclins

Cell division is the fundamental process by which all living things grow and reproduce. In most organisms, the cell cycle consists of three principle steps; interphase, mitosis, and cytokinesis. Interphase, involves preparations for cell division, replication of the DNA and production of essential proteins. In mitosis, the nuclear material is divided and separates to opposite sides of the cell. Cytokinesis is the final division and fission of the cell cytoplasm to produce the daughter cells.

The entry and exit of a cell from mitosis is regulated by the synthesis and destruction of a family of activating proteins called cyclins. Cyclins act by binding to and activating a group of cyclin-dependent protein kinases (Cdks) which then phosphorylate and activate selected proteins involved in the mitotic process. Several types of cyclins exist. (Ciechanover, A. (1994) Cell 79:13-21.) Two principle types are mitotic cyclin, or cyclin B, which controls entry of the cell into mitosis, and G1 cyclin, which controls events that drive the cell out of mitosis.

Signal Complex Scaffolding Proteins

Ceretain proteins in intracellular signaling pathways serve to link or cluster other proteins involved in the signaling cascade. A conserved protein domain called the PDZ domain has been identified in various membrane-associated signaling proteins. This domain has been implicated in receptor and ion channel clustering and in the targeting of multiprotein signaling complexes to specialized functional regions of the cytosolic face of the plasma membrane. (For a review of PDZ domain-containing proteins, see Ponting, C.P. et al. (1997) Bioessays 19:469-479.) A large proportion of PDZ domains are found in the eukaryotic MAGUK (membrane-associated guanylate kinase) protein family, members of which bind to the intracellular domains of receptors and channels. However, PDZ domains are also found in diverse membrane-localized proteins such as protein tyrosine phosphatases, serine/threonine kinases, G-protein cofactors, and synapse-associated proteins such as syntrophins and neuronal nitric oxide synthase (nNOS). Generally, about one to three PDZ domains are found in a given protein, although up to nine PDZ domains have been identified in a single protein.

**Membrane Transport Molecules**

SEQ ID NO:13 encodes, for example, a membrane transport molecule.

The plasma membrane acts as a barrier to most molecules. Transport between the cytoplasm and the extracellular environment, and between the cytoplasm and lumenal spaces of cellular organelles requires specific transport proteins. Each transport protein carries a particular class of molecule, such as ions, sugars, or amino acids, and often is specific to a certain molecular species of the class. A variety of human inherited diseases are caused by a mutation in a transport protein. For example, cystinuria is an inherited disease that results from the inability to transport cystine, the disulfide-linked dimer of cysteine, from the urine into the blood. Accumulation of cystine in the urine leads to the formation of cystine stones in the kidneys.

Transport proteins are multi-pass transmembrane proteins, which either actively transport molecules across the membrane or passively allow them to cross. Active transport involves directional pumping of a solute across the membrane, usually against an electrochemical gradient. Active transport is tightly coupled to a source of metabolic energy, such as ATP hydrolysis or an electrochemically favorable ion gradient. Passive transport involves the movement of a solute down its electrochemical gradient. Transport proteins can be further classified as either carrier proteins or channel proteins. Carrier proteins, which can function in active or passive transport, bind to a specific solute to be transported and undergo a conformational change which transfers the bound solute across the membrane. Channel proteins, which only function in passive transport, form hydrophilic pores across the membrane. When the pores open, specific solutes, such as inorganic ions, pass through the membrane and down the electrochemical gradient of the solute.

Carrier proteins which transport a single solute from one side of the membrane to the other are called uniporters. In contrast, coupled transporters link the transfer of one solute with simultaneous or sequential transfer of a second solute, either in the same direction (symport) or in the opposite direction (antiport). For example, intestinal and kidney epithelium contains a variety of symporter systems driven by the sodium gradient that exists across the plasma membrane. Sodium moves into the cell down its electrochemical gradient and brings the solute into the cell with it. The sodium gradient that provides the driving force for solute uptake is maintained by the ubiquitous $Na^+/K^+$ ATPase. Sodium-coupled transporters include the mammalian glucose transporter (SGLT1), iodide transporter (NIS), and multivitamin transporter (SMVT). All three transporters have twelve putative transmembrane segments, extracellular glycosylation sites, and cytoplasmically-oriented N- and C-termini. NIS plays a crucial role in the evaluation, diagnosis, and treatment of various thyroid pathologies because it is the molecular basis for radioiodide thyroid-imaging techniques and for specific targeting of radioisotopes to the thyroid gland (Levy, O. et al. (1997) Proc. Natl. Acad. Sci. USA 94:5568-5573). SMVT is expressed in the intestinal mucosa, kidney, and placenta, and is

implicated in the transport of the water-soluble vitamins, e.g., biotin and pantothenate (Prasad, P.D. et al. (1998) J. Biol. Chem. 273:7501-7506).

Transporters play a major role in the regulation of pH, excretion of drugs, and the cellular $K^+/Na^+$ balance. Monocarboxylate anion transporters are proton-coupled symporters with a broad substrate specificity that includes L-lactate, pyruvate, and the ketone bodies acetate, acetoacetate, and beta-hydroxybutyrate. At least seven isoforms have been identified to date. The isoforms are predicted to have twelve transmembrane (TM) helical domains with a large intracellular loop between TM6 and TM7, and play a critical role in maintaining intracellular pH by removing the protons that are produced stoichiometrically with lactate during glycolysis. The best characterized H(+)-monocarboxylate transporter is that of the erythrocyte membrane, which transports L-lactate and a wide range of other aliphatic monocarboxylates. Other cells possess H(+)-linked monocarboxylate transporters with differing substrate and inhibitor selectivities. In particular, cardiac muscle and tumor cells have transporters that differ in their $K_m$ values for certain substrates, including stereoselectivity for L- over D-lactate, and in their sensitivity to inhibitors. There are Na(+)-monocarboxylate cotransporters on the luminal surface of intestinal and kidney epithelia, which allow the uptake of lactate, pyruvate, and ketone bodies in these tissues. In addition, there are specific and selective transporters for organic cations and organic anions in organs including the kidney, intestine and liver. Organic anion transporters are selective for hydrophobic, charged molecules with electron-attracting side groups. Organic cation transporters, such as the ammonium transporter, mediate the secretion of a variety of drugs and endogenous metabolites, and contribute to the maintenance of intercellular pH. (Poole, R.C. and A.P. Halestrap (1993) Am. J. Physiol. 264:C761-C782; Price, N.T. et al. (1998) Biochem. J. 329:321-328; and Martinelle, K. and I. Haggstrom (1993) J. Biotechnol. 30: 339-350.)

The largest and most diverse family of transport proteins known is the ATP-binding cassette (ABC) transporters. As a family, ABC transporters can transport substances that differ markedly in chemical structure and size, ranging from small molecules such as ions, sugars, amino acids, peptides, and phospholipids, to lipopeptides, large proteins, and complex hydrophobic drugs. ABC proteins consist of four modules: two nucleotide-binding domains (NBD), which hydrolyze ATP to supply the energy required for transport, and two membrane-spanning domains (MSD), each containing six putative transmembrane segments. These four modules may be encoded by a single gene, as is the case for the cystic fibrosis transmembrane regulator (CFTR), or by separate genes. When encoded by separate genes, each gene product contains a single NBD and MSD. These "half-molecules" form homo- and heterodimers, such as Tap1 and Tap2, the endoplasmic reticulum-based major histocompatibility (MHC) peptide transport system. Several genetic diseases are attributed to defects in ABC transporters, such as the following diseases and their corresponding proteins: cystic fibrosis (CFTR, an ion channel), adrenoleukodystrophy (adrenoleukodystrophy protein, ALDP), Zellweger

syndrome (peroxisomal membrane protein-70, PMP70), and hyperinsulinemic hypoglycemia (sulfonylurea receptor, SUR). Overexpression of the multidrug resistance (MDR) protein, another ABC transporter, in human cancer cells makes the cells resistant to a variety of cytotoxic drugs used in chemotherapy (Taglight, D. and S. Michaelis (1998) Meth. Enzymol. 292:131-163).

Transport of fatty acids across the plasma membrane can occur by diffusion, a high capacity, low affinity process. However, under normal physiological conditions a significant fraction of fatty acid transport appears to occur via a high affinity, low capacity protein-mediated transport process. Fatty acid transport protein (FATP), an integral membrane protein with four transmembrane segments, is expressed in tissues exhibiting high levels of plasma membrane fatty acid flux, such as muscle, heart, and adipose. Expression of FATP is upregulated in 3T3-L1 cells during adipose conversion, and expression in COS7 fibroblasts elevates uptake of long-chain fatty acids (Hui, T.Y. et al. (1998) J. Biol. Chem. 273:27420-27429).

Ion Channels

The electrical potential of a cell is generated and maintained by controlling the movement of ions across the plasma membrane. The movement of ions requires ion channels, which form an ion-selective pore within the membrane. There are two basic types of ion channels, ion transporters and gated ion channels. Ion transporters utilize the energy obtained from ATP hydrolysis to actively transport an ion against the ion's concentration gradient. Gated ion channels allow passive flow of an ion down the ion's electrochemical gradient under restricted conditions. Together, these types of ion channels generate, maintain, and utilize an electrochemical gradient that is used in 1) electrical impulse conduction down the axon of a nerve cell, 2) transport of molecules into cells against concentration gradients, 3) initiation of muscle contraction, and 4) endocrine cell secretion.

Ion transporters generate and maintain the resting electrical potential of a cell. Utilizing the energy derived from ATP hydrolysis, they transport ions against the ion's concentration gradient. These transmembrane ATPases are divided into three families. The phosphorylated (P) class ion transporters, including $Na^+$-$K^+$ ATPase, $Ca^{2+}$-ATPase, and $H^+$-ATPase, are activated by a phosphorylation event. P-class ion transporters are responsible for maintaining resting potential distributions such that cytosolic concentrations of $Na^+$ and $Ca^{2+}$ are low and cytosolic concentration of $K^+$ is high. The vacuolar (V) class of ion transporters includes $H^+$ pumps on intracellular organelles, such as lysosomes and Golgi. V-class ion transporters are responsible for generating the low pH within the lumen of these organelles that is required for function. The coupling factor (F) class consists of $H^+$ pumps in the mitochondria. F-class ion transporters utilize a proton gradient to generate ATP from ADP and inorganic phosphate ($P_i$).

The resting potential of the cell is utilized in many processes involving carrier proteins and gated ion channels. Carrier proteins utilize the resting potential to transport molecules into and out of

the cell. Amino acid and glucose transport into many cells is linked to sodium ion co-transport (symport) so that the movement of $Na^+$ down an electrochemical gradient drives transport of the other molecule up a concentration gradient. Similarly, cardiac muscle links transfer of $Ca^{2+}$ out of the cell with transport of $Na^+$ into the cell (antiport).

Ion channels share common structural and mechanistic themes. The channel consists of four or five subunits or protein monomers that are arranged like a barrel in the plasma membrane. Each subunit typically consists of six potential transmembrane segments (S1, S2, S3, S4, S5, and S6). The center of the barrel forms a pore lined by α-helices or β-strands. The side chains of the amino acid residues comprising the α-helices or β-strands establish the charge (cation or anion) selectivity of the channel. The degree of selectivity, or what specific ions are allowed to pass through the channel, depends on the diameter of the narrowest part of the pore.

Gated ion channels control ion flow by regulating the opening and closing of pores. These channels are categorized according to the manner of regulating the gating function. Mechanically-gated channels open pores in response to mechanical stress, voltage-gated channels open pores in response to changes in membrane potential, and ligand-gated channels open pores in the presence of a specific ion, nucleotide, or neurotransmitter.

Voltage-gated $Na^+$ and $K^+$ channels are necessary for the function of electrically excitable cells, such as nerve and muscle cells. Action potentials, which lead to neurotransmitter release and muscle contraction, arise from large, transient changes in the permeability of the membrane to $Na^+$ and $K^+$ ions. Depolarization of the membrane beyond the threshold level opens voltage-gated $Na^+$ channels. Sodium ions flow into the cell, further depolarizing the membrane and opening more voltage-gated $Na^+$ channels, which propagates the depolarization down the length of the cell. Depolarization also opens voltage-gated potassium channels. Consequently, potassium ions flow outward, which leads to repolarization of the membrane. Voltage-gated channels utilize charged residues in the fourth transmembrane segment (S4) to sense voltage change. The open state lasts only about 1 millisecond, at which time the channel spontaneously converts into an inactive state that cannot be opened irrespective of the membrane potential. Inactivation is mediated by the channel's N-terminus, which acts as a plug that closes the pore. The transition from an inactive to a closed state requires a return to resting potential.

Voltage-gated $Na^+$ channels are heterotrimeric complexes composed of a 260 kDa pore forming α subunit that associates with two smaller auxiliary subunits, β1 and β2. The β2 subunit is an integral membrane glycoprotein that contains an extracellular Ig domain, and its association with α and β1 subunits correlates with increased functional expression of the channel, a change in its gating properties, and an increase in whole cell capacitance due to an increase in membrane surface area. (Isom, L.L. et al. (1995) Cell 83:433-442.)

Voltage-gated $Ca^{2+}$ channels are involved in presynaptic neurotransmitter release, and heart and skeletal muscle contraction. The voltage-gated $Ca^{2+}$ channels from skeletal muscle (L-type) and brain (N-type) have been purified, and though their functions differ dramatically, they have similar subunit compositions. The channels are composed of three subunits. The $\alpha_1$ subunit forms the membrane pore and voltage sensor, while the $\alpha_2\delta$ and $\beta$ subunits modulate the voltage-dependence, gating properties, and the current amplitude of the channel. These subunits are encoded by at least six $\alpha_1$, one $\alpha_2\delta$, and four $\beta$ genes. A fourth subunit, $\gamma$, has been identified in skeletal muscle. (Walker, D. et al. (1998) J. Biol. Chem. 273:2361-2367; and Jay, S.D. et al. (1990) Science 248:490-492.)

Chloride channels are necessary in endocrine secretion and in regulation of cytosolic and organelle pH. In secretory epithelial cells, $Cl^-$ enters the cell across a basolateral membrane through an $Na^+$, $K^+/Cl^-$ cotransporter, accumulating in the cell above its electrochemical equilibrium concentration. Secretion of $Cl^-$ from the apical surface, in response to hormonal stimulation, leads to flow of $Na^+$ and water into the secretory lumen. The cystic fibrosis transmembrane conductance regulator (CFTR) is a chloride channel encoded by the gene for cystic fibrosis, a common fatal genetic disorder in humans. Loss of CFTR function decreases transepithelial water secretion and, as a result, the layers of mucus that coat the respiratory tree, pancreatic ducts, and intestine are dehydrated and difficult to clear. The resulting blockage of these sites leads to pancreatic insufficiency, "meconium ileus", and devastating "chronic obstructive pulmonary disease" (Al-Awqati, Q. et al. (1992) J. Exp. Biol. 172:245-266).

Many intracellular organelles contain $H^+$-ATPase pumps that generate transmembrane pH and electrochemical differences by moving protons from the cytosol to the organelle lumen. If the membrane of the organelle is permeable to other ions, then the electrochemical gradient can be abrogated without affecting the pH differential. In fact, removal of the electrochemical barrier allows more $H^+$ to be pumped across the membrane, increasing the pH differential. $Cl^-$ is the sole counterion of $H^+$ translocation in a number of organelles, including chromaffin granules, Golgi vesicles, lysosomes, and endosomes. Functions that require a low vacuolar pH include uptake of small molecules such as biogenic amines in chromaffin granules, processing of vacuolar constituents such as pro-hormones by proteolytic enzymes, and protein degradation in lysosomes (Al-Awqati, supra).

Ligand-gated channels open their pores when an extracellular or intracellular mediator binds to the channel. Neurotransmitter-gated channels are channels that open when a neurotransmitter binds to their extracellular domain. These channels exist in the postsynaptic membrane of nerve or muscle cells. There are two types of neurotransmitter-gated channels. Sodium channels open in response to excitatory neurotransmitters, such as acetylcholine, glutamate, and serotonin. This opening causes an influx of $Na^+$ and produces the initial localized depolarization that activates the voltage-gated channels and starts the action potential. Chloride channels open in response to inhibitory neurotransmitters, such as $\gamma$-aminobutyric acid (GABA) and glycine, leading to hyperpolarization of the membrane and the

subsequent generation of an action potential.

Ligand-gated channels can be regulated by intracellular second messengers. Calcium-activated $K^+$ channels are gated by internal calcium ions. In nerve cells, an influx of calcium during depolarization opens $K^+$ channels to modulate the magnitude of the action potential (Ishi, T.M. et al. (1997) Proc. Natl. Acad. Sci. USA 94:11651-11656). Cyclic nucleotide-gated (CNG) channels are gated by cytosolic cyclic nucleotides. The best examples of these are the cAMP-gated $Na^+$ channels involved in olfaction and the cGMP-gated cation channels involved in vision. Both systems involve ligand-mediated activation of a G-protein coupled receptor which then alters the level of cyclic nucleotide within the cell.

Ion channels are expressed in a number of tissues where they are implicated in a variety of processes. CNG channels, while abundantly expressed in photoreceptor and olfactory sensory cells, are also found in kidney, lung, pineal, retinal ganglion cells, testis, aorta, and brain. Calcium-activated $K^+$ channels may be responsible for the vasodilatory effects of bradykinin in the kidney and for shunting excess $K^+$ from brain capillary endothelial cells into the blood. They are also implicated in repolarizing granulocytes after agonist-stimulated depolarization (Ishi, supra). Ion channels have been the target for many drug therapies. Neurotransmitter-gated channels have been targeted in therapies for treatment of insomnia, anxiety, depression, and schizophrenia. Voltage-gated channels have been targeted in therapies for arrhythmia, ischemic stroke, head trauma, and neurodegenerative disease (Taylor, C.P. and L.S. Narasimhan (1997) Adv. Pharmacol. 39:47-98).

Disease Correlation

The etiology of numerous human diseases and disorders can be attributed to defects in the transport of molecules across membranes. Defects in the trafficking of membrane-bound transporters and ion channels are associated with several disorders, e.g. cystic fibrosis, glucose-galactose malabsorption syndrome, hypercholesterolemia, von Gierke disease, and certain forms of diabetes mellitus. Single-gene defect diseases resulting in an inability to transport small molecules across membranes include, e.g., cystinuria, iminoglycinuria, Hartup disease, and Fanconi disease (van't Hoff, W.G. (1996) Exp. Nephrol. 4:253-262; Talente, G.M. et al. (1994) Ann. Intern. Med. 120:218-226; and Chillon, M. et al. (1995) New Engl. J. Med. 332:1475-1480).

**Protein Modification and Maintenance Molecules**

The cellular processes regulating modification and maintenance of protein molecules coordinate their conformation, stabilization, and degradation. Each of these processes is mediated by key enzymes or proteins such as proteases, protease inhibitors, transferases, isomerases, and molecular chaperones.

Proteases

Proteases cleave proteins and peptides at the peptide bond that forms the backbone of the peptide and protein chain. Proteolytic processing is essential to cell growth, differentiation, remodeling, and homeostasis as well as inflammation and immune response. Typical protein half-lives range from hours to a few days, so that within all living cells, precursor proteins are being cleaved to their active form, signal sequences proteolytically removed from targeted proteins, and aged or defective proteins degraded by proteolysis. Proteases function in bacterial, parasitic, and viral invasion and replication within a host. Four principal categories of mammalian proteases have been identified based on active site structure, mechanism of action, and overall three-dimensional structure. (Beynon, R.J. and J.S. Bond (1994) Proteolytic Enzymes: A Practical Approach, Oxford University Press, New York NY, pp. 1-5).

The serine proteases (SPs) have a serine residue, usually within a conserved sequence, in an active site composed of the serine, an aspartate, and a histidine residue. SPs include the digestive enzymes trypsin and chymotrypsin, components of the complement cascade and the blood-clotting cascade, and enzymes that control extracellular protein degradation. The main SP sub-families are trypases, which cleave after arginine or lysine; aspartases, which cleave after aspartate; chymases, which cleave after phenylalanine or leucine; metases, which cleavage after methionine; and serases which cleave after serine. Enterokinase, the initiator of intestinal digestion, is a serine protease found in the intestinal brush border, where it cleaves the acidic propeptide from trypsinogen to yield active trypsin (Kitamoto, Y. et al. (1994) Proc. Natl. Acad. Sci. USA 91:7588-7592). Prolylcarboxypeptidase, a lysosomal serine peptidase that cleaves peptides such as angiotensin II and III and [des-Arg9] bradykinin, shares sequence homology with members of both the serine carboxypeptidase and prolylendopeptidase families (Tan, F. et al. (1993) J. Biol. Chem. 268:16631-16638).

Cysteine proteases (CPs) have a cysteine as the major catalytic residue at an active site where catalysis proceeds via an intermediate thiol ester and is facilitated by adjacent histidine and aspartic acid residues. CPs are involved in diverse cellular processes ranging from the processing of precursor proteins to intracellular degradation. Mammalian CPs include lysosomal cathepsins and cytosolic calcium activated proteases, calpains. CPs are produced by monocytes, macrophages and other cells of the immune system which migrate to sites of inflammation and secrete molecules involved in tissue repair. Overabundance of these repair molecules plays a role in certain disorders. In autoimmune diseases such as rheumatoid arthritis, secretion of the cysteine peptidase cathepsin C degrades collagen, laminin, elastin and other structural proteins found in the extracellular matrix of bones.

Aspartic proteases are members of the cathepsin family of lysosomal proteases and include pepsin A, gastricsin, chymosin, renin, and cathepsins D and E. Aspartic proteases have a pair of

aspartic acid residues in the active site, and are most active in the pH 2 - 3 range, in which one of the aspartate residues is ionized, the other un-ionized. Aspartic proteases include bacterial penicillopepsin, mammalian pepsin, renin, chymosin, and certain fungal proteases. Abnormal regulation and expression of cathepsins is evident in various inflammatory disease states. In cells isolated from inflamed synovia, the mRNA for stromelysin, cytokines, TIMP-1, cathepsin, gelatinase, and other molecules is preferentially expressed. Expression of cathepsins L and D is elevated in synovial tissues from patients with rheumatoid arthritis and osteoarthritis. Cathepsin L expression may also contribute to the influx of mononuclear cells which exacerbates the destruction of the rheumatoid synovium. (Keyszer, G.M. (1995) Arthritis Rheum. 38:976-984.) The increased expression and differential regulation of the cathepsins are linked to the metastatic potential of a variety of cancers and as such are of therapeutic and prognostic interest (Chambers, A.F. et al. (1993) Crit. Rev. Oncog. 4:95-114).

Metalloproteases have active sites that include two glutamic acid residues and one histidine residue that serve as binding sites for zinc. Carboxypeptidases A and B are the principal mammalian metalloproteases. Both are exoproteases of similar structure and active sites. Carboxypeptidase A, like chymotrypsin, prefers C-terminal aromatic and aliphatic side chains of hydrophobic nature, whereas carboxypeptidase B is directed toward basic arginine and lysine residues. Glycoprotease (GCP), or O-sialoglycoprotein endopeptidase, is a metallopeptidase which specifically cleaves O-sialoglycoproteins such as glycophorin A. Another metallopeptidase, placental leucine aminopeptidase (P-LAP) degrades several peptide hormones such as oxytocin and vasopressin, suggesting a role in maintaining homeostasis during pregnancy, and is expressed in several tissues (Rogi, T. et al. (1996) J. Biol. Chem. 271:56-61).

Ubiquitin proteases are associated with the ubiquitin conjugation system (UCS), a major pathway for the degradation of cellular proteins in eukaryotic cells and some bacteria. The UCS mediates the elimination of abnormal proteins and regulates the half-lives of important regulatory proteins that control cellular processes such as gene transcription and cell cycle progression. In the UCS pathway, proteins targeted for degradation are conjugated to a ubiquitin, a small heat stable protein. The ubiquitinated protein is then recognized and degraded by proteasome, a large, multisubunit proteolytic enzyme complex, and ubiquitin is released for reutilization by ubiquitin protease. The UCS is implicated in the degradation of mitotic cyclic kinases, oncoproteins, tumor suppressor genes such as p53, viral proteins, cell surface receptors associated with signal transduction, transcriptional regulators, and mutated or damaged proteins (Ciechanover, A. (1994) Cell 79:13-21). A murine proto-oncogene, Unp, encodes a nuclear ubiquitin protease whose overexpression leads to oncogenic transformation of NIH3T3 cells, and the human homolog of this gene is consistently elevated in small cell tumors and adenocarcinomas of the lung (Gray, D.A.

(1995) Oncogene 10:2179-2183).

Signal Peptidases

The mechanism for the translocation process into the endoplasmic reticulum (ER) involves the recognition of an N-terminal signal peptide on the elongating protein. The signal peptide directs the protein and attached ribosome to a receptor on the ER membrane. The polypeptide chain passes through a pore in the ER membrane into the lumen while the N-terminal signal peptide remains attached at the membrane surface. The process is completed when signal peptidase located inside the ER cleaves the signal peptide from the protein and releases the protein into the lumen.

Protease Inhibitors

Protease inhibitors and other regulators of protease activity control the activity and effects of proteases. Protease inhibitors have been shown to control pathogenesis in animal models of proteolytic disorders (Murphy, G. (1991) Agents Actions Suppl. 35:69-76). Low levels of the cystatins, low molecular weight inhibitors of the cysteine proteases, correlate with malignant progression of tumors. (Calkins, C. et al (1995) Biol. Biochem. Hoppe Seyler 376:71-80). Serpins are inhibitors of mammalian plasma serine proteases. Many serpins serve to regulate the blood clotting cascade and/or the complement cascade in mammals. Sp32 is a positive regulator of the mammalian acrosomal protease, acrosin, that binds the proenzyme, proacrosin, and thereby aides in packaging the enzyme into the acrosomal matrix (Baba, T. et al. (1994) J. Biol. Chem. 269:10133-10140). The Kunitz family of serine protease inhibitors are characterized by one or more "Kunitz domains" containing a series of cysteine residues that are regularly spaced over approximately 50 amino acid residues and form three intrachain disulfide bonds. Members of this family include aprotinin, tissue factor pathway inhibitor (TFPI-1 and TFPI-2), inter-α-trypsin inhibitor, and bikunin. (Marlor, C.W. et al. (1997) J. Biol. Chem. 272:12202-12208.) Members of this family are potent inhibitors (in the nanomolar range) against serine proteases such as kallikrein and plasmin. Aprotinin has clinical utility in reduction of perioperative blood loss.

A major portion of all proteins synthesized in eukaryotic cells are synthesized on the cytosolic surface of the endoplasmic reticulum (ER). Before these immature proteins are distributed to other organelles in the cell or are secreted, they must be transported into the interior lumen of the ER where post-translational modifications are performed. These modifications include protein folding and the formation of disulfide bonds, and N-linked glycosylations.

Protein Isomerases

Protein folding in the ER is aided by two principal types of protein isomerases, protein disulfide isomerase (PDI), and peptidyl-prolyl isomerase (PPI). PDI catalyzes the oxidation of free sulfhydryl groups in cysteine residues to form intramolecular disulfide bonds in proteins. PPI, an enzyme that catalyzes the isomerization of certain proline imidic bonds in oligopeptides and proteins,

is considered to govern one of the rate limiting steps in the folding of many proteins to their final functional conformation. The cyclophilins represent a major class of PPI that was originally identified as the major receptor for the immunosuppressive drug cyclosporin A (Handschumacher, R.E. et al. (1984) Science 226: 544-547).

Protein Glycosylation

The glycosylation of most soluble secreted and membrane-bound proteins by oligosaccharides linked to asparagine residues in proteins is also performed in the ER. This reaction is catalyzed by a membrane-bound enzyme, oligosaccharyl transferase. Although the exact purpose of this "N-linked" glycosylation is unknown, the presence of oligosaccharides tends to make a glycoprotein resistant to protease digestion. In addition, oligosaccharides attached to cell-surface proteins called selectins are known to function in cell-cell adhesion processes (Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing Co., New York NY, p.608). "O-linked" glycosylation of proteins also occurs in the ER by the addition of N-acetylgalactosamine to the hydroxyl group of a serine or threonine residue followed by the sequential addition of other sugar residues to the first. This process is catalysed by a series of glycosyltransferases each specific for a particular donor sugar nucleotide and acceptor molecule (Lodish, H. et al. (1995) Molecular Cell Biology, W.H. Freeman and Co., New York NY, pp.700-708). In many cases, both N- and O-linked oligosaccharides appear to be required for the secretion of proteins or the movement of plasma membrane glycoproteins to the cell surface.

An additional glycosylation mechanism operates in the ER specifically to target lysosomal enzymes to lysosomes and prevent their secretion. Lysosomal enzymes in the ER receive an N-linked oligosaccharide, like plasma membrane and secreted proteins, but are then phosphorylated on one or two mannose residues. The phosphorylation of mannose residues occurs in two steps, the first step being the addition of an N-acetylglucosamine phosphate residue by N-acetylglucosamine phosphotransferase, and the second the removal of the N-acetylglucosamine group by phosphodiesterase. The phosphorylated mannose residue then targets the lysosomal enzyme to a mannose 6-phosphate receptor which transports it to a lysosome vesicle (Lodish, supra, pp. 708-711).

Chaperones

Molecular chaperones are proteins that aid in the proper folding of immature proteins and refolding of improperly folded ones, the assembly of protein subunits, and in the transport of unfolded proteins across membranes. Chaperones are also called heat-shock proteins (hsp) because of their tendency to be expressed in dramatically increased amounts following brief exposure of cells to elevated temperatures. This latter property most likely reflects their need in the refolding of proteins that have become denatured by the high temperatures. Chaperones may be divided into several classes according to their location, function, and molecular weight, and include hsp60, TCP1, hsp70,

hsp40 (also called DnaJ), and hsp90. For example, hsp90 binds to steroid hormone receptors, represses transcription in the absence of the ligand, and provides proper folding of the ligand-binding domain of the receptor in the presence of the hormone (Burston, S.G. and A.R. Clarke (1995) Essays Biochem. 29:125-136). Hsp60 and hsp70 chaperones aid in the transport and folding of newly synthesized proteins. Hsp70 acts early in protein folding, binding a newly synthesized protein before it leaves the ribosome and transporting the protein to the mitochondria or ER before releasing the folded protein. Hsp60, along with hsp10, binds misfolded proteins and gives them the opportunity to refold correctly. All chaperones share an affinity for hydrophobic patches on incompletely folded proteins and the ability to hydrolyze ATP. The energy of ATP hydrolysis is used to release the hsp-bound protein in its properly folded state (Alberts, supra, pp 214, 571-572).

**Nucleic Acid Synthesis and Modification Molecules**

SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, and SEQ ID NO:20 encode, for example, nucleic acid synthesis and modification molecules.

Polymerases

DNA and RNA replication are critical processes for cell replication and function. DNA and RNA replication are mediated by the enzymes DNA and RNA polymerase, respectively, by a "templating" process in which the nucleotide sequence of a DNA or RNA strand is copied by complementary base-pairing into a complementary nucleic acid sequence of either DNA or RNA. However, there are fundamental differences between the two processes.

DNA polymerase catalyzes the stepwise addition of a deoxyribonucleotide to the 3'-OH end of a polynucleotide strand (the primer strand) that is paired to a second (template) strand. The new DNA strand therefore grows in the 5' to 3' direction (Alberts, B. et al. (1994) The Molecular Biology of the Cell, Garland Publishing Inc., New York NY, pp. 251-254). The substrates for the polymerization reaction are the corresponding deoxynucleotide triphosphates which must base-pair with the correct nucleotide on the template strand in order to be recognized by the polymerase. Because DNA exists as a double-stranded helix, each of the two strands may serve as a template for the formation of a new complementary strand. Each of the two daughter cells of the dividing cell therefore inherits a new DNA double helix containing one old and one new strand. Thus, DNA is said to be replicated "semiconservatively" by DNA polymerase. In addition to the synthesis of new DNA, DNA polymerase is also involved in the repair of damaged DNA as discussed below under "Ligases."

In contrast to DNA polymerase, RNA polymerase uses a DNA template strand to "transcribe" DNA into RNA using ribonucleotide triphosphates as substrates. Like DNA polymerization, RNA polymerization proceeds in a 5' to 3' direction by addition of a ribonucleoside monophosphate to the 3'-OH end of a growing RNA chain. DNA transcription generates messenger RNAs (mRNA) that carry

information for protein synthesis, as well as the transfer, ribosomal, and other RNAs that have structural or catalytic functions. In eukaryotes, three discrete RNA polymerases synthesize the three different types of RNA (Alberts, supra, pp. 367-368). RNA polymerase I makes the large ribosomal RNAs, RNA polymerase II makes the mRNAs that will be translated into proteins, and RNA polymerase III makes a variety of small, stable RNAs, including 5S ribosomal RNA and the transfer RNAs (tRNA). In all cases, RNA synthesis is initiated by binding of the RNA polymerase to a promoter region on the DNA and synthesis begins at a start site within the promoter. Synthesis is completed at a broad, general stop or termination region in the DNA where both the polymerase and the completed RNA chain are released.

Ligases

DNA repair is the process by which accidental base changes, such as those produced by oxidative damage, hydrolytic attack, or uncontrolled methylation of DNA are corrected before replication or transcription of the DNA can occur. Because of the efficiency of the DNA repair process, fewer than one in one thousand accidental base changes causes a mutation (Alberts, supra, pp. 245-249). The three steps common to most types of DNA repair are (1) excision of the damaged or altered base or nucleotide by DNA nucleases, leaving a gap; (2) insertion of the correct nucleotide in this gap by DNA polymerase using the complementary strand as the template; and (3) sealing the break left between the inserted nucleotide(s) and the existing DNA strand by DNA ligase. In the last reaction, DNA ligase uses the energy from ATP hydrolysis to activate the 5' end of the broken phosphodiester bond before forming the new bond with the 3'-OH of the DNA strand. In Bloom's syndrome, an inherited human disease, individuals are partially deficient in DNA ligation and consequently have an increased incidence of cancer (Alberts, supra, p. 247).

Nucleases

Nucleases comprise both enzymes that hydrolyze DNA (DNase) and RNA (RNase). They serve different purposes in nucleic acid metabolism. Nucleases hydrolyze the phosphodiester bonds between adjacent nucleotides either at internal positions (endonucleases) or at the terminal 3' or 5' nucleotide positions (exonucleases). A DNA exonuclease activity in DNA polymerase, for example, serves to remove improperly paired nucleotides attached to the 3'-OH end of the growing DNA strand by the polymerase and thereby serves a "proofreading" function. As mentioned above, DNA endonuclease activity is involved in the excision step of the DNA repair process.

RNases also serve a variety of functions. For example, RNase P is a ribonucleoprotein enzyme which cleaves the 5' end of pre-tRNAs as part of their maturation process. RNase H digests the RNA strand of an RNA/DNA hybrid. Such hybrids occur in cells invaded by retroviruses, and RNase H is an important enzyme in the retroviral replication cycle. Pancreatic RNase secreted by the pancreas into the intestine hydrolyzes RNA present in ingested foods. RNase activity in serum and cell extracts is

elevated in a variety of cancers and infectious diseases (Schein, C.H. (1997) Nat. Biotechnol. 15:529-536). Regulation of RNase activity is being investigated as a means to control tumor angiogenesis, allergic reactions, viral infection and replication, and fungal infections.

Methylases

Methylation of specific nucleotides occurs in both DNA and RNA, and serves different functions in the two macromolecules. Methylation of cytosine residues to form 5-methyl cytosine in DNA occurs specifically at CG sequences which are base-paired with one another in the DNA double-helix. This pattern of methylation is passed from generation to generation during DNA replication by an enzyme called "maintenance methylase" that acts preferentially on those CG sequences that are base-paired with a CG sequence that is already methylated. Such methylation appears to distinguish active from inactive genes by preventing the binding of regulatory proteins that "turn on" the gene, but permit the binding of proteins that inactivate the gene (Alberts, supra, pp. 448-451). In RNA metabolism, "tRNA methylase" produces one of several nucleotide modifications in tRNA that affect the conformation and base-pairing of the molecule and facilitate the recognition of the appropriate mRNA codons by specific tRNAs. The primary methylation pattern is the dimethylation of guanine residues to form N,N-dimethyl guanine.

Helicases and Single-Stranded Binding Proteins

Helicases are enzymes that destabilize and unwind double helix structures in both DNA and RNA. Since DNA replication occurs more or less simultaneously on both strands, the two strands must first separate to generate a replication "fork" for DNA polymerase to act on. Two types of replication proteins contribute to this process, DNA helicases and single-stranded binding proteins. DNA helicases hydrolyze ATP and use the energy of hydrolysis to separate the DNA strands. Single-stranded binding proteins (SSBs) then bind to the exposed DNA strands without covering the bases, thereby temporarily stabilizing them for templating by the DNA polymerase (Alberts, supra, pp. 255-256).

RNA helicases also alter and regulate RNA conformation and secondary structure. Like the DNA helicases, RNA helicases utilize energy derived from ATP hydrolysis to destabilize and unwind RNA duplexes. The most well-characterized and ubiquitous family of RNA helicases is the DEAD-box family, so named for the conserved B-type ATP-binding motif which is diagnostic of proteins in this family. Over 40 DEAD-box helicases have been identified in organisms as diverse as bacteria, insects, yeast, amphibians, mammals, and plants. DEAD-box helicases function in diverse processes such as translation initiation, splicing, ribosome assembly, and RNA editing, transport, and stability. Some DEAD-box helicases play tissue- and stage-specific roles in spermatogenesis and embryogenesis. Overexpression of the DEAD-box 1 protein (DDX1) may play a role in the progression of neuroblastoma (Nb) and retinoblastoma (Rb) tumors (Godbout, R. et al. (1998) J. Biol. Chem. 273:21161-21168). These observations suggest that DDX1 may promote or enhance tumor

progression by altering the normal secondary structure and expression levels of RNA in cancer cells. Other DEAD-box helicases have been implicated either directly or indirectly in tumorigenesis (Discussed in Godbout, supra). For example, murine p68 is mutated in ultraviolet light-induced tumors, and human DDX6 is located at a chromosomal breakpoint associated with B-cell lymphoma. Similarly, a chimeric protein comprised of DDX10 and NUP98, a nucleoporin protein, may be involved in the pathogenesis of certain myeloid malignancies.

Topoisomerases

Besides the need to separate DNA strands prior to replication, the two strands must be "unwound" from one another prior to their separation by DNA helicases. This function is performed by proteins known as DNA topoisomerases. DNA topoisomerase effectively acts as a reversible nuclease that hydrolyzes a phosphodiesterase bond in a DNA strand, permitting the two strands to rotate freely about one another to remove the strain of the helix, and then rejoins the original phosphodiester bond between the two strands. Two types of DNA topoisomerase exist, types I and II. DNA Topoisomerase I causes a single-strand break in a DNA helix to allow the rotation of the two strands of the helix about the remaining phosphodiester bond in the opposite strand. DNA topoisomerase II causes a transient break in both strands of a DNA helix where two double helices cross over one another. This type of topoisomerase can efficiently separate two interlocked DNA circles (Alberts, supra, pp.260-262). Type II topoisomerases are largely confined to proliferating cells in eukaryotes, such as cancer cells. For this reason they are targets for anticancer drugs. Topoisomerase II has been implicated in multi-drug resistance (MDR) as it appears to aid in the repair of DNA damage inflicted by DNA binding agents such as doxorubicin and vincristine.

Recombinases

Genetic recombination is the process of rearranging DNA sequences within an organism's genome to provide genetic variation for the organism in response to changes in the environment. DNA recombination allows variation in the particular combination of genes present in an individual's genome, as well as the timing and level of expression of these genes (see Alberts, supra, pp. 263-273). Two broad classes of genetic recombination are commonly recognized, general recombination and site-specific recombination. General recombination involves genetic exchange between any homologous pair of DNA sequences usually located on two copies of the same chromosome. The process is aided by enzymes called recombinases that "nick" one strand of a DNA duplex more or less randomly and permit exchange with the complementary strand of another duplex. The process does not normally change the arrangement of genes on a chromosome. In site-specific recombination, the recombinase recognizes specific nucleotide sequences present in one or both of the recombining molecules. Base-pairing is not involved in this form of recombination and therefore does not require DNA homology between the recombining molecules. Unlike general recombination, this form of recombination can

alter the relative positions of nucleotide sequences in chromosomes.

Splicing Factors

Various proteins are necessary for processing of transcribed RNAs in the nucleus. Pre-mRNA processing steps include capping at the 5' end with methylguanosine, polyadenylating the 3' end, and splicing to remove introns. The primary RNA transcript from DNA is a faithful copy of the gene containing both exon and intron sequences, and the latter sequences must be cut out of the RNA transcript to produce an mRNA that codes for a protein. This "splicing" of the mRNA sequence takes place in the nucleus with the aid of a large, multicomponent ribonucleoprotein complex known as a spliceosome. The spliceosomal complex is composed of five small nuclear ribonucleoprotein particles (snRNPs) designated U1, U2, U4, U5, and U6, and a number of additional proteins. Each snRNP contains a single species of snRNA and about ten proteins. The RNA components of some snRNPs recognize and base pair with intron consensus sequences. The protein components mediate spliceosome assembly and the splicing reaction. Autoantibodies to snRNP proteins are found in the blood of patients with systemic lupus erythematosus (Stryer, L. (1995) Biochemistry, W.H. Freeman and Company, New York NY, p. 863).

**Adhesion Molecules**

SEQ ID NO:21 and SEQ ID NO:22 encode, for example, adhesion molecules.

The surface of a cell is rich in transmembrane proteoglycans, glycoproteins, glycolipids, and receptors. These macromolecules mediate adhesion with other cells and with components of the extracellular matrix (ECM). The interaction of the cell with its surroundings profoundly influences cell shape, strength, flexibility, motility, and adhesion. These dynamic properties are intimately associated with signal transduction pathways controlling cell proliferation and differentiation, tissue construction, and embryonic development.

Cadherins

Cadherins comprise a family of calcium-dependent glycoproteins that function in mediating cell-cell adhesion in virtually all solid tissues of multicellular organisms. These proteins share multiple repeats of a cadherin-specific motif, and the repeats form the folding units of the cadherin extracellular domain. Cadherin molecules cooperate to form focal contacts, or adhesion plaques, between adjacent epithelial cells. The cadherin family includes the classical cadherins and protocadherins. Classical cadherins include the E-cadherin, N-cadherin, and P-cadherin subfamilies. E-cadherin is present on many types of epithelial cells and is especially important for embryonic development. N-cadherin is present on nerve, muscle, and lens cells and is also critical for embryonic development. P-cadherin is present on cells of the placenta and epidermis. Recent studies report that protocadherins are involved in a variety of cell-cell interactions (Suzuki, S.T. (1996) J. Cell Sci. 109:2609-2611). The intracellular

anchorage of cadherins is regulated by their dynamic association with catenins, a family of cytoplasmic signal transduction proteins associated with the actin cytoskeleton. The anchorage of cadherins to the actin cytoskeleton appears to be regulated by protein tyrosine phosphorylation, and the cadherins are the target of phosphorylation-induced junctional disassembly (Aberle, H. et al. (1996) J. Cell. Biochem. 61:514-523).

Integrins

Integrins are ubiquitous transmembrane adhesion molecules that link the ECM to the internal cytoskeleton. Integrins are composed of two noncovalently associated transmembrane glycoprotein subunits called α and β. Integrins function as receptors that play a role in signal transduction. For example, binding of integrin to its extracellular ligand may stimulate changes in intracellular calcium levels or protein kinase activity (Sjaastad, M.D. and W.J. Nelson (1997) BioEssays 19:47-55). At least ten cell surface receptors of the integrin family recognize the ECM component fibronectin, which is involved in many different biological processes including cell migration and embryogenesis (Johansson, S. et al. (1997) Front. Biosci. 2:D126-D146).

Lectins

Lectins comprise a ubiquitous family of extracellular glycoproteins which bind cell surface carbohydrates specifically and reversibly, resulting in the agglutination of cells (reviewed in Drickamer, K. and M.E. Taylor (1993) Annu. Rev. Cell Biol. 9:237-264). This function is particularly important for activation of the immune response. Lectins mediate the agglutination and mitogenic stimulation of lymphocytes at sites of inflammation (Lasky, L.A. (1991) J. Cell. Biochem. 45:139-146; Paietta, E. et al. (1989) J. Immunol. 143:2850-2857).

Lectins are further classified into subfamilies based on carbohydrate-binding specificity and other criteria. The galectin subfamily, in particular, includes lectins that bind β-galactoside carbohydrate moieties in a thiol-dependent manner (reviewed in Hadari, Y.R. et al. (1998) J. Biol. Chem. 270:3447-3453). Galectins are widely expressed and developmentally regulated. Because all galectins lack an N-terminal signal peptide, it is suggested that galectins are externalized through an atypical secretory mechanism. Two classes of galectins have been defined based on molecular weight and oligomerization properties. Small galectins form homodimers and are about 14 to 16 kilodaltons in mass, while large galectins are monomeric and about 29-37 kilodaltons.

Galectins contain a characteristic carbohydrate recognition domain (CRD). The CRD is about 140 amino acids and contains several stretches of about 1 - 10 amino acids which are highly conserved among all galectins. A particular 6-amino acid motif within the CRD contains conserved tryptophan and arginine residues which are critical for carbohydrate binding. The CRD of some galectins also contains cysteine residues which may be important for disulfide bond formation. Secondary structure predictions indicate that the CRD forms several β-sheets.

Galectins play a number of roles in diseases and conditions associated with cell-cell and cell-matrix interactions. For example, certain galectins associate with sites of inflammation and bind to cell surface immunoglobulin E molecules. In addition, galectins may play an important role in cancer metastasis. Galectin overexpression is correlated with the metastatic potential of cancers in humans and mice. Moreover, anti-galectin antibodies inhibit processes associated with cell transformation, such as cell aggregation and anchorage-independent growth (See, for example, Su, Z.-Z. et al. (1996) Proc. Natl. Acad. Sci. USA 93:7252-7257).

Selectins

Selectins, or LEC-CAMs, comprise a specialized lectin subfamily involved primarily in inflammation and leukocyte adhesion (Reviewed in Lasky, supra). Selectins mediate the recruitment of leukocytes from the circulation to sites of acute inflammation and are expressed on the surface of vascular endothelial cells in response to cytokine signaling. Selectins bind to specific ligands on the leukocyte cell membrane and enable the leukocyte to adhere to and migrate along the endothelial surface. Binding of selectin to its ligand leads to polarized rearrangement of the actin cytoskeleton and stimulates signal transduction within the leukocyte (Brenner, B. et al. (1997) Biochem. Biophys. Res. Commun. 231:802-807; Hidari, K.I. et al. (1997) J. Biol. Chem. 272:28750-28756). Members of the selectin family possess three characteristic motifs: a lectin or carbohydrate recognition domain; an epidermal growth factor-like domain; and a variable number of short consensus repeats (scr or "sushi" repeats) which are also present in complement regulatory proteins. The selectins include lymphocyte adhesion molecule-1 (Lam-1 or L-selectin), endothelial leukocyte adhesion molecule-1 (ELAM-1 or E-selectin), and granule membrane protein-140 (GMP-140 or P-selectin) (Johnston, G.I. et al. (1989) Cell 56:1033-1044).

**Antigen Recognition Molecules**

All vertebrates have developed sophisticated and complex immune systems that provide protection from viral, bacterial, fungal, and parasitic infections. A key feature of the immune system is its ability to distinguish foreign molecules, or antigens, from "self" molecules. This ability is mediated primarily by secreted and transmembrane proteins expressed by leukocytes (white blood cells) such as lymphocytes, granulocytes, and monocytes. Most of these proteins belong to the immunoglobulin (Ig) superfamily, members of which contain one or more repeats of a conserved structural domain. This Ig domain is comprised of antiparallel β sheets joined by a disulfide bond in an arrangement called the Ig fold. Members of the Ig superfamily include T-cell receptors, major histocompatibility (MHC) proteins, antibodies, and immune cell-specific surface markers such as CD4, CD8, and CD28.

MHC proteins are cell surface markers that bind to and present foreign antigens to T cells.

MHC molecules are classified as either class I or class II. Class I MHC molecules (MHC I) are expressed on the surface of almost all cells and are involved in the presentation of antigen to cytotoxic T cells. For example, a cell infected with virus will degrade intracellular viral proteins and express the protein fragments bound to MHC I molecules on the cell surface. The MHC I/antigen complex is recognized by cytotoxic T-cells which destroy the infected cell and the virus within. Class II MHC molecules are expressed primarily on specialized antigen-presenting cells of the immune system, such as B-cells and macrophages. These cells ingest foreign proteins from the extracellular fluid and express MHC II/antigen complex on the cell surface. This complex activates helper T-cells, which then secrete cytokines and other factors that stimulate the immune response. MHC molecules also play an important role in organ rejection following transplantation. Rejection occurs when the recipient's T-cells respond to foreign MHC molecules on the transplanted organ in the same way as to self MHC molecules bound to foreign antigen. (Reviewed in Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing, New York NY, pp. 1229-1246.)

Antibodies, or immunoglobulins, are either expressed on the surface of B-cells or secreted by B-cells into the circulation. Antibodies bind and neutralize foreign antigens in the blood and other extracellular fluids. The prototypical antibody is a tetramer consisting of two identical heavy polypeptide chains (H-chains) and two identical light polypeptide chains (L-chains) interlinked by disulfide bonds. This arrangement confers the characteristic Y-shape to antibody molecules. Antibodies are classified based on their H-chain composition. The five antibody classes, IgA, IgD, IgE, IgG and IgM, are defined by the $\alpha$, $\delta$, $\epsilon$, $\gamma$, and $\mu$ H-chain types. There are two types of L-chains, $\kappa$ and $\lambda$, either of which may associate as a pair with any H-chain pair. IgG, the most common class of antibody found in the circulation, is tetrameric, while the other classes of antibodies are generally variants or multimers of this basic structure.

H-chains and L-chains each contain an N-terminal variable region and a C-terminal constant region. The constant region consists of about 110 amino acids in L-chains and about 330 or 440 amino acids in H-chains. The amino acid sequence of the constant region is nearly identical among H- or L-chains of a particular class. The variable region consists of about 110 amino acids in both H- and L-chains. However, the amino acid sequence of the variable region differs among H- or L-chains of a particular class. Within each H- or L-chain variable region are three hypervariable regions of extensive sequence diversity, each consisting of about 5 to 10 amino acids. In the antibody molecule, the H- and L-chain hypervariable regions come together to form the antigen recognition site. (Reviewed in Alberts, supra, pp. 1206-1213 and 1216-1217.)

Both H-chains and L-chains contain repeated Ig domains. For example, a typical H-chain contains four Ig domains, three of which occur within the constant region and one of which occurs within the variable region and contributes to the formation of the antigen recognition site. Likewise,

a typical L-chain contains two Ig domains, one of which occurs within the constant region and one of which occurs within the variable region.

The immune system is capable of recognizing and responding to any foreign molecule that enters the body. Therefore, the immune system must be armed with a full repertoire of antibodies against all potential antigens. Such antibody diversity is generated by somatic rearrangement of gene segments encoding variable and constant regions. These gene segments are joined together by site-specific recombination which occurs between highly conserved DNA sequences that flank each gene segment. Because there are hundreds of different gene segments, millions of unique genes can be generated combinatorially. In addition, imprecise joining of these segments and an unusually high rate of somatic mutation within these segments further contribute to the generation of a diverse antibody population.

T-cell receptors are both structurally and functionally related to antibodies. (Reviewed in Alberts, supra, pp. 1228-1229.) T-cell receptors are cell surface proteins that bind foreign antigens and mediate diverse aspects of the immune response. A typical T-cell receptor is a heterodimer comprised of two disulfide-linked polypeptide chains called α and β. Each chain is about 280 amino acids in length and contains one variable region and one constant region. Each variable or constant region folds into an Ig domain. The variable regions from the α and β chains come together in the heterodimer to form the antigen recognition site. T-cell receptor diversity is generated by somatic rearrangement of gene segments encoding the α and β chains. T-cell receptors recognize small peptide antigens that are expressed on the surface of antigen-presenting cells and pathogen-infected cells. These peptide antigens are presented on the cell surface in association with major histocompatibility proteins which provide the proper context for antigen recognition.

**Secreted and Extracellular Matrix Molecules**

SEQ ID NO:25 encodes, for example, a secreted/extracellular matrix molecule.

Protein secretion is essential for cellular function. Protein secretion is mediated by a signal peptide located at the amino terminus of the protein to be secreted. The signal peptide is comprised of about ten to twenty hydrophobic amino acids which target the nascent protein from the ribosome to the endoplasmic reticulum (ER). Proteins targeted to the ER may either proceed through the secretory pathway or remain in any of the secretory organelles such as the ER, Golgi apparatus, or lysosomes. Proteins that transit through the secretory pathway are either secreted into the extracellular space or retained in the plasma membrane. Secreted proteins are often synthesized as inactive precursors that are activated by post-translational processing events during transit through the secretory pathway. Such events include glycosylation, proteolysis, and removal of the signal peptide by a signal peptidase. Other events that may occur during protein transport include chaperone-dependent unfolding and

folding of the nascent protein and interaction of the protein with a receptor or pore complex. Examples of secreted proteins with amino terminal signal peptides include receptors, extracellular matrix molecules, cytokines, hormones, growth and differentiation factors, neuropeptides, vasomediators, ion channels, transporters/pumps, and proteases. (Reviewed in Alberts, B. et al. (1994) Molecular Biology of The Cell, Garland Publishing, New York NY, pp. 557-560, 582-592.)

The extracellular matrix (ECM) is a complex network of glycoproteins, polysaccharides, proteoglycans, and other macromolecules that are secreted from the cell into the extracellular space. The ECM remains in close association with the cell surface and provides a supportive meshwork that profoundly influences cell shape, motility, strength, flexibility, and adhesion. In fact, adhesion of a cell to its surrounding matrix is required for cell survival except in the case of metastatic tumor cells, which have overcome the need for cell-ECM anchorage. This phenomenon suggests that the ECM plays a critical role in the molecular mechanisms of growth control and metastasis. (Reviewed in Ruoslahti, E. (1996) Sci. Am. 275:72-77.) Furthermore, the ECM determines the structure and physical properties of connective tissue and is particularly important for morphogenesis and other processes associated with embryonic development and pattern formation.

The collagens comprise a family of ECM proteins that provide structure to bone, teeth, skin, ligaments, tendons, cartilage, blood vessels, and basement membranes. Multiple collagen proteins have been identified. Three collagen molecules fold together in a triple helix stabilized by interchain disulfide bonds. Bundles of these triple helices then associate to form fibrils. Collagen primary structure consists of hundreds of (Gly-X-Y) repeats where about a third of the X and Y residues are Pro. Glycines are crucial to helix formation as the bulkier amino acid sidechains cannot fold into the triple helical conformation. Because of these strict sequence requirements, mutations in collagen genes have severe consequences. Osteogenesis imperfecta patients have brittle bones that fracture easily; in severe cases patients die in utero or at birth. Ehlers-Danlos syndrome patients have hyperelastic skin, hypermobile joints, and susceptibility to aortic and intestinal rupture. Chondrodysplasia patients have short stature and ocular disorders. Alport syndrome patients have hematuria, sensorineural deafness, and eye lens deformation. (Isselbacher, K.J. et al. (1994) Harrison's Principles of Internal Medicine, McGraw-Hill, Inc., New York NY, pp. 2105-2117; and Creighton, T.E. (1984) Proteins, Structures and Molecular Principles, W.H. Freeman and Company, New York NY, pp. 191-197.)

Elastin and related proteins confer elasticity to tissues such as skin, blood vessels, and lungs. Elastin is a highly hydrophobic protein of about 750 amino acids that is rich in proline and glycine residues. Elastin molecules are highly cross-linked, forming an extensive extracellular network of fibers and sheets. Elastin fibers are surrounded by a sheath of microfibrils which are composed of a number of glycoproteins, including fibrillin. Mutations in the gene encoding fibrillin are responsible for Marfan's syndrome, a genetic disorder characterized by defects in connective tissue. In severe cases,

the aortas of afflicted individuals are prone to rupture. (Reviewed in Alberts, supra, pp. 984-986.)

Fibronectin is a large ECM glycoprotein found in all vertebrates. Fibronectin exists as a dimer of two subunits, each containing about 2,500 amino acids. Each subunit folds into a rod-like structure containing multiple domains. The domains each contain multiple repeated modules, the most common of which is the type III fibronectin repeat. The type III fibronectin repeat is about 90 amino acids in length and is also found in other ECM proteins and in some plasma membrane and cytoplasmic proteins. Furthermore, some type III fibronectin repeats contain a characteristic tripeptide consisting of Arginine-Glycine-Aspartic acid (RGD). The RGD sequence is recognized by the integrin family of cell surface receptors and is also found in other ECM proteins. Disruption of both copies of the gene encoding fibronectin causes early embryonic lethality in mice. The mutant embryos display extensive morphological defects, including defects in the formation of the notochord, somites, heart, blood vessels, neural tube, and extraembryonic structures. (Reviewed in Alberts, supra, pp. 986-987.)

Laminin is a major glycoprotein component of the basal lamina which underlies and supports epithelial cell sheets. Laminin is one of the first ECM proteins synthesized in the developing embryo. Laminin is an 850 kilodalton protein composed of three polypeptide chains joined in the shape of a cross by disulfide bonds. Laminin is especially important for angiogenesis and in particular, for guiding the formation of capillaries. (Reviewed in Alberts, supra, pp. 990-991.)

There are many other types of proteinaceous ECM components, most of which can be classified as proteoglycans. Proteoglycans are composed of unbranched polysaccharide chains (glycosaminoglycans) attached to protein cores. Common proteoglycans include aggrecan, betaglycan, decorin, perlecan, serglycin, and syndecan-1. Some of these molecules not only provide mechanical support, but also bind to extracellular signaling molecules, such as fibroblast growth factor and transforming growth factor β, suggesting a role for proteoglycans in cell-cell communication and cell growth. (Reviewed in Alberts, supra, pp. 973-978.) Likewise, the glycoproteins tenascin-C and tenascin-R are expressed in developing and lesioned neural tissue and provide stimulatory and anti-adhesive (inhibitory) properties, respectively, for axonal growth. (Faissner, A. (1997) Cell Tissue Res. 290:331-341.)

**Cytoskeletal Molecules**

SEQ ID NO:26 and SEQ ID NO:27 encode, for example, cytoskeletal molecules.

The cytoskeleton is a cytoplasmic network of protein fibers that mediate cell shape, structure, and movement. The cytoskeleton supports the cell membrane and forms tracks along which organelles and other elements move in the cytosol. The cytoskeleton is a dynamic structure that allows cells to adopt various shapes and to carry out directed movements. Major cytoskeletal fibers include the microtubules, the microfilaments, and the intermediate filaments. Motor proteins,

including myosin, dynein, and kinesin, drive movement of or along the fibers. The motor protein dynamin drives the formation of membrane vesicles. Accessory or associated proteins modify the structure or activity of the fibers while cytoskeletal membrane anchors connect the fibers to the cell membrane.

Tubulins

Microtubules, cytoskeletal fibers with a diameter of about 24 nm, have multiple roles in the cell. Bundles of microtubules form cilia and flagella, which are whip-like extensions of the cell membrane that are necessary for sweeping materials across an epithelium and for swimming of sperm, respectively. Marginal bands of microtubules in red blood cells and platelets are important for these cells' pliability. Organelles, membrane vesicles, and proteins are transported in the cell along tracks of microtubules. For example, microtubules run through nerve cell axons, allowing bi-directional transport of materials and membrane vesicles between the cell body and the nerve terminal. Failure to supply the nerve terminal with these vesicles blocks the transmission of neural signals. Microtubules are also critical to chromosomal movement during cell division. Both stable and short-lived populations of microtubules exist in the cell.

Microtubules are polymers of GTP-binding tubulin protein subunits. Each subunit is a heterodimer of α- and β- tubulin, multiple isoforms of which exist. The hydrolysis of GTP is linked to the addition of tubulin subunits at the end of a microtubule. The subunits interact head to tail to form protofilaments; the protofilaments interact side to side to form a microtubule. A microtubule is polarized, one end ringed with α-tubulin and the other with β-tubulin, and the two ends differ in their rates of assembly. Generally, each microtubule is composed of 13 protofilaments although 11 or 15 protofilament-microtubules are sometimes found. Cilia and flagella contain doublet microtubules. Microtubules grow from specialized structures known as centrosomes or microtubule-organizing centers (MTOCs). MTOCs may contain one or two centrioles, which are pinwheel arrays of triplet microtubules. The basal body, the organizing center located at the base of a cilium or flagellum, contains one centriole. Gamma tubulin present in the MTOC is important for nucleating the polymerization of α- and β- tubulin heterodimers but does not polymerize into microtubules.

Microtubule-Associated Proteins

Microtubule-associated proteins (MAPs) have roles in the assembly and stabilization of microtubules. One major family of MAPs, assembly MAPs, can be identified in neurons as well as non-neuronal cells. Assembly MAPs are responsible for cross-linking microtubules in the cytosol. These MAPs are organized into two domains: a basic microtubule-binding domain and an acidic projection domain. The projection domain is the binding site for membranes, intermediate filaments, or other microtubules. Based on sequence analysis, assembly MAPs can be further grouped into two types: Type I and Type II. Type I MAPs, which include MAP1A and MAP1B, are large, filamentous

molecules that co-purify with microtubules and are abundantly expressed in brain and testes. Type I MAPs contain several repeats of a positively-charged amino acid sequence motif that binds and neutralizes negatively charged tubulin, leading to stabilization of microtubules. MAP1A and MAP1B are each derived from a single precursor polypeptide that is subsequently proteolytically processed to generate one heavy chain and one light chain.

Another light chain, LC3, is a 16.4 kDa molecule that binds MAP1A, MAP1B, and microtubules. It is suggested that LC3 is synthesized from a source other than the MAP1A or MAP1B transcripts, and that the expression of LC3 may be important in regulating the microtubule binding activity of MAP1A and MAP1B during cell proliferation (Mann, S.S. et al. (1994) J. Biol. Chem. 269:11492-11497).

Type II MAPs, which include MAP2a, MAP2b, MAP2c, MAP4, and Tau, are characterized by three to four copies of an 18-residue sequence in the microtubule-binding domain. MAP2a, MAP2b, and MAP2c are found only in dendrites, MAP4 is found in non-neuronal cells, and Tau is found in axons and dendrites of nerve cells. Alternative splicing of the Tau mRNA leads to the existence of multiple forms of Tau protein. Tau phosphorylation is altered in neurodegenerative disorders such as Alzheimer's disease, Pick's disease, progressive supranuclear palsy, corticobasal degeneration, and familial frontotemporal dementia and Parkinsonism linked to chromosome 17. The altered Tau phosphorylation leads to a collapse of the microtubule network and the formation of intraneuronal Tau aggregates (Spillantini, M.G. and M. Goedert (1998) Trends Neurosci. 21:428-433).

The protein pericentrin is found in the MTOC and has a role in microtubule assembly.

Actins

Microfilaments, cytoskeletal filaments with a diameter of about 7-9 nm, are vital to cell locomotion, cell shape, cell adhesion, cell division, and muscle contraction. Assembly and disassembly of the microfilaments allow cells to change their morphology. Microfilaments are the polymerized form of actin, the most abundant intracellular protein in the eukaryotic cell. Human cells contain six isoforms of actin. The three α-actins are found in different kinds of muscle, nonmuscle β-actin and nonmuscle γ-actin are found in nonmuscle cells, and another γ-actin is found in intestinal smooth muscle cells. G-actin, the monomeric form of actin, polymerizes into polarized, helical F-actin filaments, accompanied by the hydrolysis of ATP to ADP. Actin filaments associate to form bundles and networks, providing a framework to support the plasma membrane and determine cell shape. These bundles and networks are connected to the cell membrane. In muscle cells, thin filaments containing actin slide past thick filaments containing the motor protein myosin during contraction. A family of actin-related proteins exist that are not part of the actin cytoskeleton, but rather associate with microtubules and dynein.

Actin-Associated Proteins

Actin-associated proteins have roles in cross-linking, severing, and stabilization of actin filaments and in sequestering actin monomers. Several of the actin-associated proteins have multiple functions. Bundles and networks of actin filaments are held together by actin cross-linking proteins. These proteins have two actin-binding sites, one for each filament. Short cross-linking proteins promote bundle formation while longer, more flexible cross-linking proteins promote network formation. Calmodulin-like calcium-binding domains in actin cross-linking proteins allow calcium regulation of cross-linking. Group I cross-linking proteins have unique actin-binding domains and include the 30 kD protein, EF-1a, fascin, and scruin. Group II cross-linking proteins have a 7,000-MW actin-binding domain and include villin and dematin. Group III cross-linking proteins have pairs of a 26,000-MW actin-binding domain and include fimbrin, spectrin, dystrophin, ABP 120, and filamin.

Severing proteins regulate the length of actin filaments by breaking them into short pieces or by blocking their ends. Severing proteins include gCAP39, severin (fragmin), gelsolin, and villin. Capping proteins can cap the ends of actin filaments, but cannot break filaments. Capping proteins include CapZ and tropomodulin. The proteins thymosin and profilin sequester actin monomers in the cytosol, allowing a pool of unpolymerized actin to exist. The actin-associated proteins tropomyosin, troponin, and caldesmon regulate muscle contraction in response to calcium.

Intermediate Filaments and Associated Proteins

Intermediate filaments (IFs) are cytoskeletal fibers with a diameter of about 10 nm, intermediate between that of microfilaments and microtubules. IFs serve structural roles in the cell, reinforcing cells and organizing cells into tissues. IFs are particularly abundant in epidermal cells and in neurons. IFs are extremely stable, and, in contrast to microfilaments and microtubules, do not function in cell motility.

Five types of IF proteins are known in mammals. Type I and Type II proteins are the acidic and basic keratins, respectively. Heterodimers of the acidic and basic keratins are the building blocks of keratin IFs. Keratins are abundant in soft epithelia such as skin and cornea, hard epithelia such as nails and hair, and in epithelia that line internal body cavities. Mutations in keratin genes lead to epithelial diseases including epidermolysis bullosa simplex, bullous congenital ichthyosiform erythroderma (epidermolytic hyperkeratosis), non-epidermolytic and epidermolytic palmoplantar keratoderma, ichthyosis bullosa of Siemens, pachyonychia congenita, and white sponge nevus. Some of these diseases result in severe skin blistering. (See, e.g., Wawersik, M. et al. (1997) J. Biol. Chem. 272:32557-32565; and Corden L.D. and W.H. McLean (1996) Exp. Dermatol. 5:297-307.)

Type III IF proteins include desmin, glial fibrillary acidic protein, vimentin, and peripherin. Desmin filaments in muscle cells link myofibrils into bundles and stabilize sarcomeres in contracting muscle. Glial fibrillary acidic protein filaments are found in the glial cells that surround neurons and

astrocytes. Vimentin filaments are found in blood vessel endothelial cells, some epithelial cells, and mesenchymal cells such as fibroblasts, and are commonly associated with microtubules. Vimentin filaments may have roles in keeping the nucleus and other organelles in place in the cell. Type IV IFs include the neurofilaments and nestin. Neurofilaments, composed of three polypeptides NF-L, NF-M, and NF-H, are frequently associated with microtubules in axons. Neurofilaments are responsible for the radial growth and diameter of an axon, and ultimately for the speed of nerve impulse transmission. Changes in phosphorylation and metabolism of neurofilaments are observed in neurodegenerative diseases including amyotrophic lateral sclerosis, Parkinson's disease, and Alzheimer's disease (Julien, J.P. and W.E. Mushynski (1998) Prog. Nucleic Acid Res. Mol. Biol. 61:1-23). Type V IFs, the lamins, are found in the nucleus where they support the nuclear membrane.

IFs have a central α-helical rod region interrupted by short nonhelical linker segments. The rod region is bracketed, in most cases, by non-helical head and tail domains. The rod regions of intermediate filament proteins associate to form a coiled-coil dimer. A highly ordered assembly process leads from the dimers to the IFs. Neither ATP nor GTP is needed for IF assembly, unlike that of microfilaments and microtubules.

IF-associated proteins (IFAPs) mediate the interactions of IFs with one another and with other cell structures. IFAPs cross-link IFs into a bundle, into a network, or to the plasma membrane, and may cross-link IFs to the microfilament and microtubule cytoskeleton. Microtubules and IFs are in particular closely associated. IFAPs include BPAG1, plakoglobin, desmoplakin I, desmoplakin II, plectin, ankyrin, filaggrin, and lamin B receptor.

Cytoskeletal-Membrane Anchors

Cytoskeletal fibers are attached to the plasma membrane by specific proteins. These attachments are important for maintaining cell shape and for muscle contraction. In erythrocytes, the spectrin-actin cytoskeleton is attached to cell membrane by three proteins, band 4.1, ankyrin, and adducin. Defects in this attachment result in abnormally shaped cells which are more rapidly degraded by the spleen, leading to anemia. In platelets, the spectrin-actin cytoskeleton is also linked to the membrane by ankyrin; a second actin network is anchored to the membrane by filamin. In muscle cells the protein dystrophin links actin filaments to the plasma membrane; mutations in the dystrophin gene lead to Duchenne muscular dystrophy. In adherens junctions and adhesion plaques the peripheral membrane proteins α-actinin and vinculin attach actin filaments to the cell membrane.

IFs are also attached to membranes by cytoskeletal-membrane anchors. The nuclear lamina is attached to the inner surface of the nuclear membrane by the lamin B receptor. Vimentin IFs are attached to the plasma membrane by ankyrin and plectin. Desmosome and hemidesmosome membrane junctions hold together epithelial cells of organs and skin. These membrane junctions allow shear forces to be distributed across the entire epithelial cell layer, thus providing strength and

rigidity to the epithelium. IFs in epithelial cells are attached to the desmosome by plakoglobin and desmoplakins. The proteins that link IFs to hemidesmosomes are not known. Desmin IFs surround the sarcomere in muscle and are linked to the plasma membrane by paranemin, synemin, and ankyrin.

Myosin-related Motor Proteins

Myosins are actin-activated ATPases, found in eukaryotic cells, that couple hydrolysis of ATP with motion. Myosin provides the motor function for muscle contraction and intracellular movements such as phagocytosis and rearrangement of cell contents during mitotic cell division (cytokinesis). The contractile unit of skeletal muscle, termed the sarcomere, consists of highly ordered arrays of thin actin-containing filaments and thick myosin-containing filaments. Crossbridges form between the thick and thin filaments, and the ATP-dependent movement of myosin heads within the thick filaments pulls the thin filaments, shortening the sarcomere and thus the muscle fiber.

Myosins are composed of one or two heavy chains and associated light chains. Myosin heavy chains contain an amino-terminal motor or head domain, a neck that is the site of light-chain binding, and a carboxy-terminal tail domain. The tail domains may associate to form an α-helical coiled coil. Conventional myosins, such as those found in muscle tissue, are composed of two myosin heavy-chain subunits, each associated with two light-chain subunits that bind at the neck region and play a regulatory role. Unconventional myosins, believed to function in intracellular motion, may contain either one or two heavy chains and associated light chains. There is evidence for about 25 myosin heavy chain genes in vertebrates, more than half of them unconventional.

Dynein-related Motor Proteins

Dyneins are (-) end-directed motor proteins which act on microtubules. Two classes of dyneins, cytosolic and axonemal, have been identified. Cytosolic dyneins are responsible for translocation of materials along cytoplasmic microtubules, for example, transport from the nerve terminal to the cell body and transport of endocytic vesicles to lysosomes. Cytoplasmic dyneins are also reported to play a role in mitosis. Axonemal dyneins are responsible for the beating of flagella and cilia. Dynein on one microtubule doublet walks along the adjacent microtubule doublet. This sliding force produces bending forces that cause the flagellum or cilium to beat. Dyneins have a native mass between 1000 and 2000 kDa and contain either two or three force-producing heads driven by the hydrolysis of ATP. The heads are linked via stalks to a basal domain which is composed of a highly variable number of accessory intermediate and light chains.

Kinesin-related Motor Proteins

Kinesins are (+) end-directed motor proteins which act on microtubules. The prototypical kinesin molecule is involved in the transport of membrane-bound vesicles and organelles. This function is particularly important for axonal transport in neurons. Kinesin is also important in all cell types for the transport of vesicles from the Golgi complex to the endoplasmic reticulum. This role is critical for

maintaining the identity and functionality of these secretory organelles.

Kinesins define a ubiquitous, conserved family of over 50 proteins that can be classified into at least 8 subfamilies based on primary amino acid sequence, domain structure, velocity of movement, and cellular function. (Reviewed in Moore, J.D. and S.A. Endow (1996) Bioessays 18:207-219; and Hoyt, A.M. (1994) Curr. Opin. Cell Biol. 6:63-68.) The prototypical kinesin molecule is a heterotetramer comprised of two heavy polypeptide chains (KHCs) and two light polypeptide chains (KLCs). The KHC subunits are typically referred to as "kinesin." KHC is about 1000 amino acids in length, and KLC is about 550 amino acids in length. Two KHCs dimerize to form a rod-shaped molecule with three distinct regions of secondary structure. At one end of the molecule is a globular motor domain that functions in ATP hydrolysis and microtubule binding. Kinesin motor domains are highly conserved and share over 70% identity. Beyond the motor domain is an α-helical coiled-coil region which mediates dimerization. At the other end of the molecule is a fan-shaped tail that associates with molecular cargo. The tail is formed by the interaction of the KHC C-termini with the two KLCs.

Members of the more divergent subfamilies of kinesins are called kinesin-related proteins (KRPs), many of which function during mitosis in eukaryotes (Hoyt, supra). Some KRPs are required for assembly of the mitotic spindle. In vivo and in vitro analyses suggest that these KRPs exert force on microtubules that comprise the mitotic spindle, resulting in the separation of spindle poles. Phosphorylation of KRP is required for this activity. Failure to assemble the mitotic spindle results in abortive mitosis and chromosomal aneuploidy, the latter condition being characteristic of cancer cells. In addition, a unique KRP, centromere protein E, localizes to the kinetochore of human mitotic chromosomes and may play a role in their segregation to opposite spindle poles.

Dynamin-related Motor Proteins

Dynamin is a large GTPase motor protein that functions as a "molecular pinchase," generating a mechanochemical force used to sever membranes. This activity is important in forming clathrin-coated vesicles from coated pits in endocytosis and in the biogenesis of synaptic vesicles in neurons. Binding of dynamin to a membrane leads to dynamin's self-assembly into spirals that may act to constrict a flat membrane surface into a tubule. GTP hydrolysis induces a change in conformation of the dynamin polymer that pinches the membrane tubule, leading to severing of the membrane tubule and formation of a membrane vesicle. Release of GDP and inorganic phosphate leads to dynamin disassembly. Following disassembly the dynamin may either dissociate from the membrane or remain associated to the vesicle and be transported to another region of the cell. Three homologous dynamin genes have been discovered, in addition to several dynamin-related proteins. Conserved dynamin regions are the N-terminal GTP-binding domain, a central pleckstrin homology domain that binds membranes, a central coiled-coil region that may activate dynamin's GTPase activity, and a C-terminal proline-rich domain that contains several motifs that bind SH3 domains on

other proteins. Some dynamin-related proteins do not contain the pleckstrin homology domain or the proline-rich domain. (See McNiven, M.A. (1998) Cell 94:151-154; Scaife, R.M. and R.L. Margolis (1997) Cell. Signal. 9:395-401.)

The cytoskeleton is reviewed in Lodish, H. et al. (1995) Molecular Cell Biology, Scientific American Books, New York NY.

**Ribosomal Molecules**

SEQ ID NO:30 and SEQ ID NO:31 encode, for example, ribosomal molecules.

Ribosomal RNAs (rRNAs) are assembled, along with ribosomal proteins, into ribosomes, which are cytoplasmic particles that translate messenger RNA into polypeptides. The eukaryotic ribosome is composed of a 60S (large) subunit and a 40S (small) subunit, which together form the 80S ribosome. In addition to the 18S, 28S, 5S, and 5.8S rRNAs, the ribosome also contains more than fifty proteins. The ribosomal proteins have a prefix which denotes the subunit to which they belong, either L (large) or S (small). Ribosomal protein activities include binding rRNA and organizing the conformation of the junctions between rRNA helices (Woodson, S.A. and N.B. Leontis (1998) Curr. Opin. Struct. Biol. 8:294-300; Ramakrishnan, V. and S.W. White (1998) Trends Biochem. Sci. 23:208-212.) Three important sites are identified on the ribosome. The aminoacyl-tRNA site (A site) is where charged tRNAs (with the exception of the initiator-tRNA) bind on arrival at the ribosome. The peptidyl-tRNA site (P site) is where new peptide bonds are formed, as well as where the initiator tRNA binds. The exit site (E site) is where deacylated tRNAs bind prior to their release from the ribosome. (The ribosome is reviewed in Stryer, L. (1995) Biochemistry W.H. Freeman and Company, New York NY, pp. 888-908; and Lodish, H. et al. (1995) Molecular Cell Biology Scientific American Books, New York NY. pp. 119-138.)

**Chromatin Molecules**

The nuclear DNA of eukaryotes is organized into chromatin. Two types of chromatin are observed: euchromatin, some of which may be transcribed, and heterochromatin so densely packed that much of it is inaccessible to transcription. Chromatin packing thus serves to regulate protein expression in eukaryotes. Bacteria lack chromatin and the chromatin-packing level of gene regulation.

The fundamental unit of chromatin is the nucleosome of 200 DNA base pairs associated with two copies each of histones H2A, H2B, H3, and H4. Adjascent nucleosomes are linked by another class of histones, H1. Low molecular weight non-histone proteins called the high mobility group (HMG), associated with chromatin, may function in the unwinding of DNA and stabilization of single-stranded DNA. Chromodomain proteins function in compaction of chromatin into its transcriptionally silent heterochromatin form.

During mitosis, all DNA is compacted into heterochromatin and transcription ceases. Transcription in interphase begins with the activation of a region of chromatin. Active chromatin is decondensed. Decondensation appears to be accompanied by changes in binding coefficient, phosphorylation and acetylation states of chromatin histones. HMG proteins HMG13 and HMG17 selectively bind activated chromatin. Topoisomerases remove superhelical tension on DNA. The activated region decondenses, allowing gene regulatory proteins and transcription factors to assemble on the DNA.

Patterns of chromatin structure can be stably inherited, producing heritable patterns of gene expression. In mammals, one of the two X chromosomes in each female cell is inactivated by condensation to heterochromatin during zygote development. The inactive state of this chromosome is inherited, so that adult females are mosaics of clusters of paternal-X and maternal-X clonal cell groups. The condensed X chromosome is reactivated in meiosis.

Chromatin is associated with disorders of protein expression such as thalassemia, a genetic anemia resulting from the removal of the locus control region (LCR) required for decondensation of the globin gene locus.

For a review of chromatin structure and function see Alberts, B. et al. (1994) Molecular Cell Biology, third edition, Garland Publishing, Inc., New York NY, pp. 351-354, 433-439.

**Electron Transfer Associated Molecules**

SEQ ID NO:23 and SEQ ID NO:24 encode, for example, electron transfer associated molecules.

Electron carriers such as cytochromes accept electrons from NADH or $FADH_2$ and donate them to other electron carriers. Most electron-transferring proteins, except ubiquinone, are prosthetic groups such as flavins, heme, FeS clusters, and copper, bound to inner membrane proteins. Adrenodoxin, for example, is an FeS protein that forms a complex with NADPH:adrenodoxin reductase and cytochrome p450. Cytochromes contain a heme prosthetic group, a porphyrin ring containing a tightly bound iron atom. Electron transfer reactions play a crucial role in cellular energy production.

Energy is produced by the oxidation of glucose and fatty acids. Glucose is initially converted to pyruvate in the cytoplasm. Fatty acids and pyruvate are transported to the mitochondria for complete oxidation to $CO_2$ coupled by enzymes to the transport of electrons from NADH and $FADH_2$ to oxygen and to the synthesis of ATP (oxidative phosphorylation) from ADP and $P_i$.

Pyruvate is transported into the mitochondria and converted to acetyl-CoA for oxidation via the citric acid cycle, involving pyruvate dehydrogenase components, dihydrolipoyl transacetylase, and dihydrolipoyl dehydrogenase. Enzymes involved in the citric acid cycle include: citrate synthetase,

aconitases, isocitrate dehydrogenase, alpha-ketoglutarate dehydrogenase complex including transsuccinylases, succinyl CoA synthetase, succinate dehydrogenase, fumarases, and malate dehydrogenase. Acetyl CoA is oxidized to $CO_2$ with concomitant formation of NADH, $FADH_2$, and GTP. In oxidative phosphorylation, the transfer of electrons from NADH and $FADH_2$ to oxygen by dehydrogenases is coupled to the synthesis of ATP from ADP and $P_i$ by the $F_0F_1$ ATPase complex in the mitochondrial inner membrane. Enzyme complexes responsible for electron transport and ATP synthesis include the $F_0F_1$ ATPase complex, ubiquinone(CoQ)-cytochrome c reductase, ubiquinone reductase, cytochrome b, cytochrome $c_1$, FeS protein, and cytochrome c oxidase.

ATP synthesis requires membrane transport enzymes including the phosphate transporter and the ATP-ADP antiport protein. The ATP-binding casette (ABC) superfamily has also been suggested as belonging to the mitochondrial transport group (Hogue, D.L. et al. (1999) J. Mol. Biol. 285:379-389). Brown fat uncoupling protein dissipates oxidative energy as heat, and may be involved the fever response to infection and trauma (Cannon, B. et al. (1998) Ann. NY Acad. Sci. 856:171-187).

Mitochondria are oval-shaped organelles comprising an outer membrane, a tightly folded inner membrane, an intermembrane space between the outer and inner membranes, and a matrix inside the inner membrane. The outer membrane contains many porin molecules that allow ions and charged molecules to enter the intermembrane space, while the inner membrane contains a variety of transport proteins that transfer only selected molecules. Mitochondria are the primary sites of energy production in cells.

Mitochondria contain a small amount of DNA. Human mitochondrial DNA encodes 13 proteins, 22 tRNAs, and 2 rRNAs. Mitochondrial-DNA encoded proteins include NADH-Q reductase, a cytochrome reductase subunit, cytochrome oxidase subunits, and ATP synthase subunits.

Electron-transfer reactions also occur outside the mitochondria in locations such as the endoplasmic reticulum, which plays a crucial role in lipid and protein biosynthesis. Cytochrome b5 is a central electron donor for various reductive reactions occurring on the cytoplasmic surface of liver endoplasmic reticulum. Cytochrome b5 has been found in Golgi, plasma, endoplasmic reticulum (ER), and microbody membranes.

For a review of mitochondrial metabolism and regulation, see Lodish, H. et al. (1995) Molecular Cell Biology, Scientific American Books, New York NY, pp. 745-797 and Stryer (1995) Biochemistry, W.H. Freeman and Co., San Francisco CA, pp 529-558, 988-989.

The majority of mitochondrial proteins are encoded by nuclear genes, are synthesized on cytosolic ribosomes, and are imported into the mitochondria. Nuclear-encoded proteins which are destined for the mitochondrial matrix typically contain positively-charged amino terminal signal sequences. Import of these preproteins from the cytoplasm requires a multisubunit protein complex in the outer membrane known as the translocase of outer mitochondrial membrane (TOM; previously

designated MOM; Pfanner, N. et al. (1996) Trends Biochem. Sci. 21:51-52) and at least three inner membrane proteins which comprise the translocase of inner mitochondrial membrane (TIM; previously designated MIM; Pfanner, supra). An inside-negative membrane potential across the inner mitochondrial membrane is also required for preprotein import. Preproteins are recognized by surface receptor components of the TOM complex and are translocated through a proteinaceous pore formed by other TOM components. Proteins targeted to the matrix are then recognized by the import machinery of the TIM complex. The import systems of the outer and inner membranes can function independently (Segui-Real, B. et al. (1993) EMBO J. 12:2211-2218).

Once precursor proteins are in the mitochondria, the leader peptide is cleaved by a signal peptidase to generate the mature protein. Most leader peptides are removed in a one step process by a protease termed mitochondrial processing peptidase (MPP) (Paces, V. et al. (1993) Proc. Natl. Acad. Sci. USA 90:5355-5358). In some cases a two-step process occurs in which MPP generates an intermediate precursor form which is cleaved by a second enzyme, mitochondrial intermediate peptidase, to generate the mature protein.

Mitochondrial dysfunction leads to impaired calcium buffering, generation of free radicals that may participate in deleterious intracellular and extracellular processes, changes in mitochondrial permeability and oxidative damage which is observed in several neurodegenerative diseases. Neurodegenerative diseases linked to mitochondrial dysfunction include some forms of Alzheimer's disease, Friedreich's ataxia, familial amyotrophic lateral sclerosis, and Huntington's disease (Beal, M.F. (1998) Biochim. Biophys. Acta 1366:211-213). The myocardium is heavily dependent on oxidative metabolism, so mitochondrial dysfunction often leads to heart disease (DiMauro, S. and M. Hirano (1998) Curr. Opin. Cardiol 13:190-197). Mitochondria are implicated in disorders of cell proliferation, since they play an important role in a cell's decision to proliferate or self-destruct through apoptosis. The oncoprotein Bcl-2, for example, promotes cell proliferation by stabilizing mitochondrial membranes so that apoptosis signals are not released (Susin, S.A. (1998) Biochim. Biophys. Acta 1366:151-165).

**Transcription Factor Molecules**

SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, SEQ ID NO:35, SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, and SEQ ID NO:43 encode, for example, transcription factor molecules.

Multicellular organisms are comprised of diverse cell types that differ dramatically both in structure and function. The identity of a cell is determined by its characteristic pattern of gene expression, and different cell types express overlapping but distinctive sets of genes throughout development. Spatial and temporal regulation of gene expression is critical for the control of cell

proliferation, cell differentiation, apoptosis, and other processes that contribute to organismal development. Furthermore, gene expression is regulated in response to extracellular signals that mediate cell-cell communication and coordinate the activities of different cell types. Appropriate gene regulation also ensures that cells function efficiently by expressing only those genes whose functions are required at a given time.

Transcriptional regulatory proteins are essential for the control of gene expression. Some of these proteins function as transcription factors that initiate, activate, repress, or terminate gene transcription. Transcription factors generally bind to the promoter, enhancer, and upstream regulatory regions of a gene in a sequence-specific manner, although some factors bind regulatory elements within or downstream of a gene's coding region. Transcription factors may bind to a specific region of DNA singly or as a complex with other accessory factors. (Reviewed in Lewin, B. (1990) Genes IV, Oxford University Press, New York NY, and Cell Press, Cambridge MA, pp. 554-570.)

The double helix structure and repeated sequences of DNA create topological and chemical features which can be recognized by transcription factors. These features are hydrogen bond donor and acceptor groups, hydrophobic patches, major and minor grooves, and regular, repeated stretches of sequence which induce distinct bends in the helix. Typically, transcription factors recognize specific DNA sequence motifs of about 20 nucleotides in length. Multiple, adjacent transcription factor-binding motifs may be required for gene regulation.

Many transcription factors incorporate DNA-binding structural motifs which comprise either α helices or ß sheets that bind to the major groove of DNA. Four well-characterized structural motifs are helix-turn-helix, zinc finger, leucine zipper, and helix-loop-helix. Proteins containing these motifs may act alone as monomers, or they may form homo- or heterodimers that interact with DNA.

The helix-turn-helix motif consists of two α helices connected at a fixed angle by a short chain of amino acids. One of the helices binds to the major groove. Helix-turn-helix motifs are exemplified by the homeobox motif which is present in homeodomain proteins. These proteins are critical for specifying the anterior-posterior body axis during development and are conserved throughout the animal kingdom. The Antennapedia and Ultrabithorax proteins of Drosophila melanogaster are prototypical homeodomain proteins (Pabo, C.O. and R.T. Sauer (1992) Annu. Rev. Biochem. 61:1053-1095).

The zinc finger motif, which binds zinc ions, generally contains tandem repeats of about 30 amino acids consisting of periodically spaced cysteine and histidine residues. Examples of this sequence pattern, designated C2H2 and C3HC4 ("RING" finger), have been described (Lewin, supra). Zinc finger proteins each contain an α helix and an antiparallel ß sheet whose proximity and conformation are maintained by the zinc ion. Contact with DNA is made by the arginine prece ding the α helix and by the second, third, and sixth residues of the α helix. Variants of the zinc finger

motif include poorly defined cysteine-rich motifs which bind zinc or other metal ions. These motifs may not contain histidine residues and are generally nonrepetitive.

The leucine zipper motif comprises a stretch of amino acids rich in leucine which can form an amphipathic α helix. This structure provides the basis for dimerization of two leucine zipper proteins. The region adjacent to the leucine zipper is usually basic, and upon protein dimerization, is optimally positioned for binding to the major groove. Proteins containing such motifs are generally referred to as bZIP transcription factors.

The helix-loop-helix motif (HLH) consists of a short α helix connected by a loop to a longer α helix. The loop is flexible and allows the two helices to fold back against each other and to bind to DNA. The transcription factor Myc contains a prototypical HLH motif.

Most transcription factors contain characteristic DNA binding motifs, and variations on the above motifs and new motifs have been and are currently being characterized (Faisst, S. and S. Meyer (1992) Nucleic Acids Res. 20:3-26).

Many neoplastic disorders in humans can be attributed to inappropriate gene expression. Malignant cell growth may result from either excessive expression of tumor promoting genes or insufficient expression of tumor suppressor genes (Cleary, M.L. (1992) Cancer Surv. 15:89-104). Chromosomal translocations may also produce chimeric loci which fuse the coding sequence of one gene with the regulatory regions of a second unrelated gene. Such an arrangement likely results in inappropriate gene transcription, potentially contributing to malignancy.

In addition, the immune system responds to infection or trauma by activating a cascade of events that coordinate the progressive selection, amplification, and mobilization of cellular defense mechanisms. A complex and balanced program of gene activation and repression is involved in this process. However, hyperactivity of the immune system as a result of improper or insufficient regulation of gene expression may result in considerable tissue or organ damage. This damage is well documented in immunological responses associated with arthritis, allergens, heart attack, stroke, and infections (Isselbacher, K.J. et al. (1996) Harrison's Principles of Internal Medicine, 13/e, McGraw Hill, Inc. and Teton Data Systems Software).

Furthermore, the generation of multicellular organisms is based upon the induction and coordination of cell differentiation at the appropriate stages of development. Central to this process is differential gene expression, which confers the distinct identities of cells and tissues throughout the body. Failure to regulate gene expression during development can result in developmental disorders. Human developmental disorders caused by mutations in zinc finger-type transcriptional regulators include: urogenenital developmental abnormalities associated with WT1; Greig cephalopolysyndactyly, Pallister-Hall syndrome, and postaxial polydactyly type A (GLI3); and Townes-Brocks syndrome, characterized by anal, renal, limb, and ear abnormalities (SALL1)

(Engelkamp, D. and V. van Heyningen (1996) Curr. Opin. Genet. Dev. 6:334-342; Kohlhase, J. et al. (1999) Am. J. Hum. Genet. 64:435-445).

**Cell Membrane Molecules**

SEQ ID NO:28 and SEQ ID NO:29 encode, for example, cell membrane molecules.

Eukaryotic cells are surrounded by plasma membranes which enclose the cell and maintain an environment inside the cell that is distinct from its surroundings. In addition, eukaryotic organisms are distinct from prokaryotes in possessing many intracellular organelle and vesicle structures. Many of the metabolic reactions which distinguish eukaryotic biochemistry from prokaryotic biochemistry take place within these structures. The plasma membrane and the membranes surrounding organelles and vesicles are composed of phosphoglycerides, fatty acids, cholesterol, phospholipids, glycolipids, proteoglycans, and proteins. These components confer identity and functionality to the membranes with which they associate.

Integral Membrane Proteins

The majority of known integral membrane proteins are transmembrane proteins (TM) which are characterized by an extracellular, a transmembrane, and an intracellular domain. TM domains are typically comprised of 15 to 25 hydrophobic amino acids which are predicted to adopt an α-helical conformation. TM proteins are classified as bitopic (Types I and II) and polytopic (Types III and IV) (Singer, S.J. (1990) Annu. Rev. Cell Biol. 6:247-296). Bitopic proteins span the membrane once while polytopic proteins contain multiple membrane-spanning segments. TM proteins function as cell-surface receptors, receptor-interacting proteins, transporters of ions or metabolites, ion channels, cell anchoring proteins, and cell type-specific surface antigens.

Many membrane proteins (MPs) contain amino acid sequence motifs that target these proteins to specific subcellular sites. Examples of these motifs include PDZ domains, KDEL, RGD, NGR, and GSL sequence motifs, von Willebrand factor A (vWFA) domains, and EGF-like domains. RGD, NGR, and GSL motif-containing peptides have been used as drug delivery agents in targeted cancer treatment of tumor vasculature (Arap, W. et al. (1998) Science 279:377-380). Furthermore, MPs may also contain amino acid sequence motifs, such as the carbohydrate recognition domain (CRD), that mediate interactions with extracellular or intracellular molecules.

G-Protein Coupled Receptors

G-protein coupled receptors (GPCR) are a superfamily of integral membrane proteins which transduce extracellular signals. GPCRs include receptors for biogenic amines, lipid mediators of inflammation, peptide hormones, and sensory signal mediators. The structure of these highly-conserved receptors consists of seven hydrophobic transmembrane regions, an extracellular N-terminus, and a cytoplasmic C-terminus. Three extracellular loops alternate with three intracellular

loops to link the seven transmembrane regions. Cysteine disulfide bridges connect the second and third extracellular loops. The most conserved regions of GPCRs are the transmembrane regions and the first two cytoplasmic loops. A conserved, acidic-Arg-aromatic residue triplet present in the second cytoplasmic loop may interact with G proteins. A GPCR consensus pattern is characteristic of most proteins belonging to this superfamily (ExPASy PROSITE document PS00237; and Watson, S. and S. Arkinstall (1994) The G-protein Linked Receptor Facts Book, Academic Press, San Diego CA, pp. 2-6). Mutations and changes in transcriptional activation of GPCR-encoding genes have been associated with neurological disorders such as schizophrenia, Parkinson's disease, Alzheimer's disease, drug addiction, and feeding disorders.

Scavenger Receptors

Macrophage scavenger receptors with broad ligand specificity may participate in the binding of low density lipoproteins (LDL) and foreign antigens. Scavenger receptors types I and II are trimeric membrane proteins with each subunit containing a small N-terminal intracellular domain, a transmembrane domain, a large extracellular domain, and a C-terminal cysteine-rich domain. The extracellular domain contains a short spacer region, an α-helical coiled-coil region, and a triple helical collagen-like region. These receptors have been shown to bind a spectrum of ligands, including chemically modified lipoproteins and albumin, polyribonucleotides, polysaccharides, phospholipids, and asbestos (Matsumoto, A. et al. (1990) Proc. Natl. Acad. Sci. USA 87:9133-9137; and Elomaa, O. et al. (1995) Cell 80:603-609). The scavenger receptors are thought to play a key role in atherogenesis by mediating uptake of modified LDL in arterial walls, and in host defense by binding bacterial endotoxins, bacteria, and protozoa.

Tetraspan Family Proteins

The transmembrane 4 superfamily (TM4SF) or tetraspan family is a multigene family encoding type III integral membrane proteins (Wright, M.D. and M.G. Tomlinson (1994) Immunol. Today 15:588-594). The TM4SF is comprised of membrane proteins which traverse the cell membrane four times. Members of the TM4SF include platelet and endothelial cell membrane proteins, melanoma-associated antigens, leukocyte surface glycoproteins, colonal carcinoma antigens, tumor-associated antigens, and surface proteins of the schistosome parasites (Jankowski, S.A. (1994) Oncogene 9:1205-1211). Members of the TM4SF share about 25-30% amino acid sequence identity with one another.

A number of TM4SF members have been implicated in signal transduction, control of cell adhesion, regulation of cell growth and proliferation, including development and oncogenesis, and cell motility, including tumor cell metastasis. Expression of TM4SF proteins is associated with a variety of tumors and the level of expression may be altered when cells are growing or activated.

Tumor Antigens

Tumor antigens are cell surface molecules that are differentially expressed in tumor cells relative to normal cells. Tumor antigens distinguish tumor cells immunologically from normal cells and provide diagnostic and therapeutic targets for human cancers (Takagi, S. et al. (1995) Int. J. Cancer 61:706-715; Liu, E. et al. (1992) Oncogene 7:1027-1032).

Leukocyte Antigens

Other types of cell surface antigens include those identified on leukocytic cells of the immune system. These antigens have been identified using systematic, monoclonal antibody (mAb)-based "shot gun" techniques. These techniques have resulted in the production of hundreds of mAbs directed against unknown cell surface leukocytic antigens. These antigens have been grouped into "clusters of differentiation" based on common immunocytochemical localization patterns in various differentiated and undifferentiated leukocytic cell types. Antigens in a given cluster are presumed to identify a single cell surface protein and are assigned a "cluster of differentiation" or "CD" designation. Some of the genes encoding proteins identified by CD antigens have been cloned and verified by standard molecular biology techniques. CD antigens have been characterized as both transmembrane proteins and cell surface proteins anchored to the plasma membrane via covalent attachment to fatty acid-containing glycolipids such as glycosylphosphatidylinositol (GPI). (Reviewed in Barclay, A.N. et al. (1995) The Leucocyte Antigen Facts Book, Academic Press, San Diego CA, pp. 17-20.)

Ion Channels

Ion channels are found in the plasma membranes of virtually every cell in the body. For example, chloride channels mediate a variety of cellular functions including regulation of membrane potentials and absorption and secretion of ions across epithelial membranes. Chloride channels also regulate the pH of organelles such as the Golgi apparatus and endosomes (see, e.g., Greger, R. (1988) Annu. Rev. Physiol. 50:111-122). Electrophysiological and pharmacological properties of chloride channels, including ion conductance, current-voltage relationships, and sensitivity to modulators, suggest that different chloride channels exist in muscles, neurons, fibroblasts, epithelial cells, and lymphocytes.

Many ion channels have sites for phosphorylation by one or more protein kinases including protein kinase A, protein kinase C, tyrosine kinase, and casein kinase II, all of which regulate ion channel activity in cells. Inappropriate phosphorylation of proteins in cells has been linked to changes in cell cycle progression and cell differentiation. Changes in the cell cycle have been linked to induction of apoptosis or cancer. Changes in cell differentiation have been linked to diseases and disorders of the reproductive system, immune system, skeletal muscle, and other organ systems.

Proton Pumps

Proton ATPases comprise a large class of membrane proteins that use the energy of ATP

hydrolysis to generate an electrochemical proton gradient across a membrane. The resultant gradient may be used to transport other ions across the membrane ($Na^+$, $K^+$, or $Cl^-$) or to maintain organelle pH. Proton ATPases are further subdivided into the mitochondrial F-ATPases, the plasma membrane ATPases, and the vacuolar ATPases. The vacuolar ATPases establish and maintain an acidic pH within various organelles involved in the processes of endocytosis and exocytosis (Mellman, I. et al. (1986) Annu. Rev. Biochem. 55:663-700).

Proton-coupled, 12 membrane-spanning domain transporters such as PEPT 1 and PEPT 2 are responsible for gastrointestinal absorption and for renal reabsorption of peptides using an electrochemical $H^+$ gradient as the driving force. Another type of peptide transporter, the TAP transporter, is a heterodimer consisting of TAP 1 and TAP 2 and is associated with antigen processing. Peptide antigens are transported across the membrane of the endoplasmic reticulum by TAP so they can be expressed on the cell surface in association with MHC molecules. Each TAP protein consists of multiple hydrophobic membrane spanning segments and a highly conserved ATP-binding cassette (Boll, M. et al. (1996) Proc. Natl. Acad. Sci. USA 93:284-289). Pathogenic microorganisms, such as herpes simplex virus, may encode inhibitors of TAP-mediated peptide transport in order to evade immune surveillance (Marusina, K. and J.J Manaco (1996) Curr. Opin. Hematol. 3:19-26).

ABC Transporters

The ATP-binding cassette (ABC) transporters, also called the "traffic ATPases", comprise a superfamily of membrane proteins that mediate transport and channel functions in prokaryotes and eukaryotes (Higgins, C.F. (1992) Annu. Rev. Cell Biol. 8:67-113). ABC proteins share a similar overall structure and significant sequence homology. All ABC proteins contain a conserved domain of approximately two hundred amino acid residues which includes one or more nucleotide binding domains. Mutations in ABC transporter genes are associated with various disorders, such as hyperbilirubinemia II/Dubin-Johnson syndrome, recessive Stargardt's disease, X-linked adrenoleukodystrophy, multidrug resistance, celiac disease, and cystic fibrosis.

Peripheral and Anchored Membrane Proteins

Some membrane proteins are not membrane-spanning but are attached to the plasma membrane via membrane anchors or interactions with integral membrane proteins. Membrane anchors are covalently joined to a protein post-translationally and include such moieties as prenyl, myristyl, and glycosylphosphatidyl inositol groups. Membrane localization of peripheral and anchored proteins is important for their function in processes such as receptor-mediated signal transduction. For example, prenylation of Ras is required for its localization to the plasma membrane and for its normal and oncogenic functions in signal transduction.

Vesicle Coat Proteins

Intercellular communication is essential for the development and survival of multicellular organisms. Cells communicate with one another through the secretion and uptake of protein signaling molecules. The uptake of proteins into the cell is achieved by the endocytic pathway, in which the interaction of extracellular signaling molecules with plasma membrane receptors results in the formation of plasma membrane-derived vesicles that enclose and transport the molecules into the cytosol. These transport vesicles fuse with and mature into endosomal and lysosomal (digestive) compartments. The secretion of proteins from the cell is achieved by exocytosis, in which molecules inside of the cell proceed through the secretory pathway. In this pathway, molecules transit from the ER to the Golgi apparatus and finally to the plasma membrane, where they are secreted from the cell.

Several steps in the transit of material along the secretory and endocytic pathways require the formation of transport vesicles. Specifically, vesicles form at the transitional endoplasmic reticulum (tER), the rim of Golgi cisternae, the face of the Trans-Golgi Network (TGN), the plasma membrane (PM), and tubular extensions of the endosomes. Vesicle formation occurs when a region of membrane buds off from the donor organelle. The membrane-bound vesicle contains proteins to be transported and is surrounded by a proteinaceous coat, the components of which are recruited from the cytosol. Two different classes of coat protein have been identified. Clathrin coats form on vesicles derived from the TGN and PM, whereas coatomer (COP) coats form on vesicles derived from the ER and Golgi. COP coats can be further classified as COPI, involved in retrograde traffic through the Golgi and from the Golgi to the ER, and COPII, involved in anterograde traffic from the ER to the Golgi (Mellman, supra).

In clathrin-based vesicle formation, adapter proteins bring vesicle cargo and coat proteins together at the surface of the budding membrane. Adapter protein-1 and -2 select cargo from the TGN and plasma membrane, respectively, based on molecular information encoded on the cytoplasmic tail of integral membrane cargo proteins. Adapter proteins also recruit clathrin to the bud site. Clathrin is a protein complex consisting of three large and three small polypeptide chains arranged in a three-legged structure called a triskelion. Multiple triskelions and other coat proteins appear to self-assemble on the membrane to form a coated pit. This assembly process may serve to deform the membrane into a budding vesicle. GTP-bound ADP-ribosylation factor (Arf) is also incorporated into the coated assembly. Another small G-protein, dynamin, forms a ring complex around the neck of the forming vesicle and may provide the mechanochemical force to seal the bud, thereby releasing the vesicle. The coated vesicle complex is then transported through the cytosol. During the transport process, Arf-bound GTP is hydrolyzed to GDP, and the coat dissociates from the transport vesicle (West, M.A. et al. (1997) J. Cell Biol. 138:1239-1254).

Vesicles which bud from the ER and the Golgi are covered with a protein coat similar to the clathrin coat of endocytic and TGN vesicles. The coat protein (COP) is assembled from cytosolic

precursor molecules at specific budding regions on the organelle. The COP coat consists of two major components, a G-protein (Arf or Sar) and coat protomer (coatomer). Coatomer is an equimolar complex of seven proteins, termed alpha-, beta-, beta'-, gamma-, delta-, epsilon- and zeta-COP. The coatomer complex binds to dilysine motifs contained on the cytoplasmic tails of integral membrane proteins. These include the KKXX retrieval motif of membrane proteins of the ER and dibasic/diphenylamine motifs of members of the p24 family. The p24 family of type I membrane proteins represent the major membrane proteins of COPI vesicles (Harter, C. and F.T. Wieland (1998) Proc. Natl. Acad. Sci. USA 95:11649-11654).

**Organelle Associated Molecules**

SEQ ID NO:44, SEQ ID NO:45, and SEQ ID NO:46 encode, for example, organelle associated molecules.

Eukaryotic cells are organized into various cellular organelles which has the effect of separating specific molecules and their functions from one another and from the cytosol. Within the cell, various membrane structures surround and define these organelles while allowing them to interact with one another and the cell environment through both active and passive transport processes. Important cell organelles include the nucleus, the Golgi apparatus, the endoplasmic reticulum, mitochondria, peroxisomes, lysosomes, endosomes, and secretory vesicles.

Nucleus

The cell nucleus contains all of the genetic information of the cell in the form of DNA, and the components and machinery necessary for replication of DNA and for transcription of DNA into RNA. (See Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing Inc., New York NY, pp. 335-399.) DNA is organized into compact structures in the nucleus by interactions with various DNA-binding proteins such as histones and non-histone chromosomal proteins. DNA-specific nucleases, DNAses, partially degrade these compacted structures prior to DNA replication or transcription. DNA replication takes place with the aid of DNA helicases which unwind the double-stranded DNA helix, and DNA polymerases that duplicate the separated DNA strands.

Transcriptional regulatory proteins are essential for the control of gene expression. Some of these proteins function as transcription factors that initiate, activate, repress, or terminate gene transcription. Transcription factors generally bind to the promoter, enhancer, and upstream regulatory regions of a gene in a sequence-specific manner, although some factors bind regulatory elements within or downstream of a gene's coding region. Transcription factors may bind to a specific region of DNA singly or as a complex with other accessory factors. (Reviewed in Lewin, B. (1990) Genes IV, Oxford University Press, New York NY, and Cell Press, Cambridge MA, pp. 554-570.)

Many transcription factors incorporate DNA-binding structural motifs which comprise either α helices or ß sheets that bind to the major groove of DNA. Four well-characterized structural motifs are helix-turn-helix, zinc finger, leucine zipper, and helix-loop-helix. Proteins containing these motifs may act alone as monomers, or they may form homo- or heterodimers that interact with DNA.

Many neoplastic disorders in humans can be attributed to inappropriate gene expression. Malignant cell growth may result from either excessive expression of tumor promoting genes or insufficient expression of tumor suppressor genes (Cleary, M.L. (1992) Cancer Surv. 15:89-104). Chromosomal translocations may also produce chimeric loci which fuse the coding sequence of one gene with the regulatory regions of a second unrelated gene. Such an arrangement likely results in inappropriate gene transcription, potentially contributing to malignancy.

In addition, the immune system responds to infection or trauma by activating a cascade of events that coordinate the progressive selection, amplification, and mobilization of cellular defense mechanisms. A complex and balanced program of gene activation and repression is involved in this process. However, hyperactivity of the immune system as a result of improper or insufficient regulation of gene expression may result in considerable tissue or organ damage. This damage is well documented in immunological responses associated with arthritis, allergens, heart attack, stroke, and infections (Isselbacher, K.J. et al. (1996) Harrison's Principles of Internal Medicine, 13/e, McGraw Hill, Inc. and Teton Data Systems Software).

Transcription of DNA into RNA also takes place in the nucleus catalyzed by RNA polymerases. Three types of RNA polymerase exist. RNA polymerase I makes large ribosomal RNAs, while RNA polymerase III makes a variety of small, stable RNAs including 5S ribosomal RNA and the transfer RNAs (tRNA). RNA polymerase II transcribes genes that will be translated into proteins. The primary transcript of RNA polymerase II is called heterogenous nuclear RNA (hnRNA), and must be further processed by splicing to remove non-coding sequences called introns. RNA splicing is mediated by small nuclear ribonucleoprotein complexes, or snRNPs, producing mature messenger RNA (mRNA) which is then transported out of the nucleus for translation into proteins.

Nucleolus

The nucleolus is a highly organized subcompartment in the nucleus that contains high concentrations of RNA and proteins and functions mainly in ribosomal RNA synthesis and assembly (Alberts, et al. supra, pp. 379-382). Ribosomal RNA (rRNA) is a structural RNA that is complexed with proteins to form ribonucleoprotein structures called ribosomes. Ribosomes provide the platform on which protein synthesis takes place.

Ribosomes are assembled in the nucleolus initially from a large, 45S rRNA combined with a variety of proteins imported from the cytoplasm, as well as smaller, 5S rRNAs. Later processing of

the immature ribosome results in formation of smaller ribosomal subunits which are transported from the nucleolus to the cytoplasm where they are assembled into functional ribosomes.

Endoplasmic Reticulum

In eukaryotes, proteins are synthesized within the endoplasmic reticulum (ER), delivered from the ER to the Golgi apparatus for post-translational processing and sorting, and transported from the Golgi to specific intracellular and extracellular destinations. Synthesis of integral membrane proteins, secreted proteins, and proteins destined for the lumen of a particular organelle occurs on the rough endoplasmic reticulum (ER). The rough ER is so named because of the rough appearance in electron micrographs imparted by the attached ribosomes on which protein synthesis proceeds. Synthesis of proteins destined for the ER actually begins in the cytosol with the synthesis of a specific signal peptide which directs the growing polypeptide and its attached ribosome to the ER membrane where the signal peptide is removed and protein synthesis is completed. Soluble proteins destined for the ER lumen, for secretion, or for transport to the lumen of other organelles pass completely into the ER lumen. Transmembrane proteins destined for the ER or for other cell membranes are translocated across the ER membrane but remain anchored in the lipid bilayer of the membrane by one or more membrane-spanning α-helical regions.

Translocated polypeptide chains destined for other organelles or for secretion also fold and assemble in the ER lumen with the aid of certain "resident" ER proteins. Protein folding in the ER is aided by two principal types of protein isomerases, protein disulfide isomerase (PDI), and peptidyl-prolyl isomerase (PPI). PDI catalyzes the oxidation of free sulfhydryl groups in cysteine residues to form intramolecular disulfide bonds in proteins. PPI, an enzyme that catalyzes the isomerization of certain proline imide bonds in oligopeptides and proteins, is considered to govern one of the rate limiting steps in the folding of many proteins to their final functional conformation. The cyclophilins represent a major class of PPI that was originally identified as the major receptor for the immunosuppressive drug cyclosporin A (Handschumacher, R.E. et al. (1984) Science 226:544-547). Molecular "chaperones" such as BiP (binding protein) in the ER recognize incorrectly folded proteins as well as proteins not yet folded into their final form and bind to them, both to prevent improper aggregation between them, and to promote proper folding.

The "N-linked" glycosylation of most soluble secreted and membrane-bound proteins by oligosacchrides linked to asparagine residues in proteins is also performed in the ER. This reaction is catalyzed by a membrane-bound enzyme, oligosaccharyl transferase.

Golgi Apparatus

The Golgi apparatus is a complex structure that lies adjacent to the ER in eukaryotic cells and serves primarily as a sorting and dispatching station for products of the ER (Alberts, et al. supra, pp. 600-610). Additional posttranslational processing, principally additional glycosylation, also occurs in

the Golgi. Indeed, the Golgi is a major site of carbohydrate synthesis, including most of the glycosaminoglycans of the extracellular matrix. N-linked oligosaccharides, added to proteins in the ER, are also further modified in the Golgi by the addition of more sugar residues to form complex N-linked oligosaccharides. "O-linked" glycosylation of proteins also occurs in the Golgi by the addition of N-acetylgalactosamine to the hydroxyl group of a serine or threonine residue followed by the sequential addition of other sugar residues to the first. This process is catalyzed by a series of glycosyltransferases each specific for a particular donor sugar nucleotide and acceptor molecule (Lodish, H. et al. (1995) Molecular Cell Biology, W.H. Freeman and Co., New York NY, pp.700-708). In many cases, both N- and O-linked oligosaccharides appear to be required for the secretion of proteins or the movement of plasma membrane glycoproteins to the cell surface.

The terminal compartment of the Golgi is the Trans-Golgi Network (TGN), where both membrane and lumenal proteins are sorted for their final destination. Transport (or secretory) vesicles destined for intracellular compartments, such as lysosomes, bud off of the TGN. Other transport vesicles bud off containing proteins destined for the plasma membrane, such as receptors, adhesion molecules, and ion channels, and secretory proteins, such as hormones, neurotransmitters, and digestive enzymes.

Vacuoles

The vacuole system is a collection of membrane bound compartments in eukaryotic cells that functions in the processes of endocytosis and exocytosis. They include phagosomes, lysosomes, endosomes, and secretory vesicles. Endocytosis is the process in cells of internalizing nutrients, solutes or small particles (pinocytosis) or large particles such as internalized receptors, viruses, bacteria, or bacterial toxins (phagocytosis). Exocytosis is the process of transporting molecules to the cell surface. It facilitates placement or localization of membrane-bound receptors or other membrane proteins and secretion of hormones, neurotransmitters, digestive enzymes, wastes, etc.

A common property of all of these vacuoles is an acidic pH environment ranging from approximately pH 4.5-5.0. This acidity is maintained by the presence of a proton ATPase that uses the energy of ATP hydrolysis to generate an electrochemical proton gradient across a membrane (Mellman, I. et al. (1986) Annu. Rev. Biochem. 55:663-700). Eukaryotic vacuolar proton ATPase (vp-ATPase) is a multimeric enzyme composed of 3-10 different subunits. One of these subunits is a highly hydrophobic polypeptide of approximately 16 kDa that is similar to the proteolipid component of vp-ATPases from eubacteria, fungi, and plant vacuoles (Mandel, M. et al. (1988) Proc. Natl. Acad. Sci. USA 85:5521-5524). The 16 kDa proteolipid component is the major subunit of the membrane portion of vp-ATPase and functions in the transport of protons across the membrane.

Lysosomes

Lysosomes are membranous vesicles containing various hydrolytic enzymes used for the

controlled intracellular digestion of macromolecules. Lysosomes contain some 40 types of enzymes including proteases, nucleases, glycosidases, lipases, phospholipases, phosphatases, and sulfatases, all of which are acid hydrolases that function at a pH of about 5. Lysosomes are surrounded by a unique membrane containing transport proteins that allow the final products of macromolecule degradation, such as sugars, amino acids, and nucleotides, to be transported to the cytosol where they may be either excreted or reutilized by the cell. A vp-ATPase, such as that described above, maintains the acidic environment necessary for hydrolytic activity (Alberts, supra, pp. 610-611).

Endosomes

Endosomes are another type of acidic vacuole that is used to transport substances from the cell surface to the interior of the cell in the process of endocytosis. Like lysosomes, endosomes have an acidic environment provided by a vp-ATPase (Alberts et al. supra, pp. 610-618). Two types of endosomes are apparent based on tracer uptake studies that distinguish their time of formation in the cell and their cellular location. Early endosomes are found near the plasma membrane and appear to function primarily in the recycling of internalized receptors back to the cell surface. Late endosomes appear later in the endocytic process close to the Golgi apparatus and the nucleus, and appear to be associated with delivery of endocytosed material to lysosomes or to the TGN where they may be recycled. Specific proteins are associated with particular transport vesicles and their target compartments that may provide selectivity in targeting vesicles to their proper compartments. A cytosolic prenylated GTP-binding protein, Rab, is one such protein. Rabs 4, 5, and 11 are associated with the early endosome, whereas Rabs 7 and 9 associate with the late endosome.

Mitochondria

Mitochondria are oval-shaped organelles comprising an outer membrane, a tightly folded inner membrane, an intermembrane space between the outer and inner membranes, and a matrix inside the inner membrane. The outer membrane contains many porin molecules that allow ions and charged molecules to enter the intermembrane space, while the inner membrane contains a variety of transport proteins that transfer only selected molecules. Mitochondria are the primary sites of energy production in cells.

Energy is produced by the oxidation of glucose and fatty acids. Glucose is initially converted to pyruvate in the cytoplasm. Fatty acids and pyruvate are transported to the mitochondria for complete oxidation to $CO_2$ coupled by enzymes to the transport of electrons from NADH and $FADH_2$ to oxygen and to the synthesis of ATP (oxidative phosphorylation) from ADP and $P_i$.

Pyruvate is transported into the mitochondria and converted to acetyl-CoA for oxidation via the citric acid cycle, involving pyruvate dehydrogenase components, dihydrolipoyl transacetylase, and dihydrolipoyl dehydrogenase. Enzymes involved in the citric acid cycle include: citrate synthetase, aconitases, isocitrate dehydrogenase, alpha-ketoglutarate dehydrogenase complex including

transsuccinylases, succinyl CoA synthetase, succinate dehydrogenase, fumarases, and malate dehydrogenase. Acetyl CoA is oxidized to $CO_2$ with concomitant formation of NADH, $FADH_2$, and GTP. In oxidative phosphorylation, the transfer of electrons from NADH and $FADH_2$ to oxygen by dehydrogenases is coupled to the synthesis of ATP from ADP and $P_i$ by the $F_0F_1$ ATPase complex in the mitochondrial inner membrane. Enzyme complexes responsible for electron transport and ATP synthesis include the $F_0F_1$ ATPase complex, ubiquinone(CoQ)-cytochrome c reductase, ubiquinone reductase, cytochrome b, cytochrome $c_1$, FeS protein, and cytochrome c oxidase.

Peroxisomes

Peroxisomes, like mitochondria, are a major site of oxygen utilization. They contain one or more enzymes, such as catalase and urate oxidase, that use molecular oxygen to remove hydrogen atoms from specific organic substrates in an oxidative reaction that produces hydrogen peroxide (Alberts, supra, pp. 574-577). Catalase oxidizes a variety of substrates including phenols, formic acid, formaldehyde, and alcohol and is important in peroxisomes of liver and kidney cells for detoxifying various toxic molecules that enter the bloodstream. Another major function of oxidative reactions in peroxisomes is the breakdown of fatty acids in a process called β oxidation. β oxidation results in shortening of the alkyl chain of fatty acids by blocks of two carbon atoms that are converted to acetyl CoA and exported to the cytosol for reuse in biosynthetic reactions.

Also like mitochondria, peroxisomes import their proteins from the cytosol using a specific signal sequence located near the C-terminus of the protein. The importance of this import process is evident in the inherited human disease Zellweger syndrome, in which a defect in importing proteins into perixosomes leads to a perixosomal deficiency resulting in severe abnormalities in the brain, liver, and kidneys, and death soon after birth. One form of this disease has been shown to be due to a mutation in the gene encoding a perixosomal integral membrane protein called peroxisome assembly factor-1.

The discovery of new human molecules for diagnostics and therapeutics satisfies a need in the art by providing new compositions which are useful in the diagnosis, study, prevention, and treatment of diseases associated with human molecules.

## SUMMARY OF THE INVENTION

The present invention relates to nucleic acid sequences comprising human diagnostic and therapeutic polynucleotides (dithp) as presented in the Sequence Listing. Some of the dithp uniquely identify genes encoding human structural, functional, and regulatory molecules.

The invention provides an isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence

identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). In one alternative, the polynucleotide comprises a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52. In another alternative, the polynucleotide comprises at least 60 contiguous nucleotides of a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The invention further provides a composition for the detection of expression of human diagnostic and therapeutic polynucleotides, comprising at least one isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d); and a detectable label.

The invention also provides a method for detecting a target polynucleotide in a sample, said target polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The method comprises a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide, and b) detecting the presence or absence of said hybridization complex, and, optionally, if present, the amount thereof. In one alternative, the probe comprises at least 30 contiguous nucleotides. In another alternative, the probe comprises at least 60 contiguous nucleotides.

The invention further provides a recombinant polynucleotide comprising a promoter sequence operably linked to an isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide

sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). In one alternative, the invention provides a cell transformed with the recombinant polynucleotide. In another alternative, the invention provides a transgenic organism comprising the recombinant polynucleotide. In a further alternative, the invention provides a method for producing a human diagnostic and therapeutic polypeptide, the method comprising a) culturing a cell under conditions suitable for expression of the human diagnostic and therapeutic polypeptide, wherein said cell is transformed with the recombinant polynucleotide, and b) recovering the human diagnostic and therapeutic polypeptide so expressed.

The invention also provides a purified human diagnostic and therapeutic polypeptide (DITHP) encoded by at least one polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52. Additionally, the invention provides an isolated antibody which specifically binds to the human diagnostic and therapeutic polypeptide. The invention further provides a method of identifying a test compound which specifically binds to the human diagnostic and therapeutic polypeptide, the method comprising the steps of a) providing a test compound; b) combining the human diagnostic and therapeutic polypeptide with the test compound for a sufficient time and under suitable conditions for binding; and c) detecting binding of the human diagnostic and therapeutic polypeptide to the test compound, thereby identifying the test compound which specifically binds the human diagnostic and therapeutic polypeptide.

The invention further provides a microarray wherein at least one element of the microarray is an isolated polynucleotide comprising at least 60 contiguous nucleotides of a polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The invention also provides a method for generating a transcript image of a sample which contains polynucleotides. The method comprises a) labeling the polynucleotides of the sample, b) contacting the elements of the microarray with the labeled polynucleotides of the sample under conditions suitable for the formation of a hybridization complex, and c) quantifying the expression of the polynucleotides in the sample.

Additionally, the invention provides a method for screening a compound for effectiveness in altering expression of a target polynucleotide, wherein said target polynucleotide comprises a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence

complementary to b); and e) an RNA equivalent of a) through d). The method comprises a) exposing a sample comprising the target polynucleotide to a compound, and b) detecting altered expression of the target polynucleotide.

The invention further provides a method for detecting a target polynucleotide in a sample for toxicity testing of a compound, said target polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The method comprises a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide, b) detecting the presence or absence of said hybridization complex, and, optionally, if present, the amount thereof, and c) comparing the presence, absence or amount of said target polynucleotide in a first biological sample and a second biological sample, wherein said first biological sample has been contacted with said compound, and said second sample is a control, whereby a change in presence, absence or amount of said target polynucleotide in said first sample, as compared with said second sample, is indicative of toxic response to said compound.

## DESCRIPTION OF THE TABLES

Table 1 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with their GenBank hits (GI Numbers), probability scores, and functional annotations corresponding to the GenBank hits.

Table 2 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with polynucleotide segments of each template sequence as defined by the indicated "start" and "stop" nucleotide positions. The reading frames of the polynucleotide segments and the Pfam hits, Pfam descriptions, and E-values corresponding to the polypeptide domains encoded by the polynucleotide segments are indicated.

Table 3 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with polynucleotide segments of each template sequence as defined by the indicated "start" and "stop" nucleotide positions. The reading frames of the polynucleotide segments are shown, and the

polypeptides encoded by the polynucleotide segments constitute either signal peptide (SP) or transmembrane (TM) domains, as indicated.

Table 4 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with component sequence identification numbers (component IDs) corresponding to each template. The component sequences, which were used to assemble the template sequences, are defined by the indicated "start" and "stop" nucleotide positions along each template.

Table 5 summarizes the bioinformatics tools which are useful for analysis of the polynucleotides of the present invention. The first column of Table 5 lists analytical tools, programs, and algorithms, the second column provides brief descriptions thereof, the third column presents appropriate references, all of which are incorporated by reference herein in their entirety, and the fourth column presents, where applicable, the scores, probability values, and other parameters used to evaluate the strength of a match between two sequences (the higher the score, the greater the homology between two sequences).

## DETAILED DESCRIPTION OF THE INVENTION

Before the nucleic acid sequences and methods are presented, it is to be understood that this invention is not limited to the particular machines, methods, and materials described. Although particular embodiments are described, machines, methods, and materials similar or equivalent to these embodiments may be used to practice the invention. The preferred machines, methods, and materials set forth are not intended to limit the scope of the invention which is limited only by the appended claims.

The singular forms "a", "an", and "the" include plural reference unless the context clearly dictates otherwise. All technical and scientific terms have the meanings commonly understood by one of ordinary skill in the art. All publications are incorporated by reference for the purpose of describing and disclosing the cell lines, vectors, and methodologies which are presented and which might be used in connection with the invention. Nothing in the specification is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

### Definitions

As used herein, the lower case "dithp" refers to a nucleic acid sequence, while the upper case "DITHP" refers to an amino acid sequence encoded by dithp. A "full-length" dithp refers to a nucleic acid sequence containing the entire coding region of a gene endogenously expressed in human tissue.

"Adjuvants" are materials such as Freund's adjuvant, mineral gels (aluminum hydroxide), and surface active substances (lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole

limpet hemocyanin, and dinitrophenol) which may be administered to increase a host's immunological response.

"Allele" refers to an alternative form of a nucleic acid sequence. Alleles result from a "mutation," a change or an alternative reading of the genetic code. Any given gene may have none, one, or many allelic forms. Mutations which give rise to alleles include deletions, additions, or substitutions of nucleotides. Each of these changes may occur alone, or in combination with the others, one or more times in a given nucleic acid sequence. The present invention encompasses allelic dithp.

"Amino acid sequence" refers to a peptide, a polypeptide, or a protein of either natural or synthetic origin. The amino acid sequence is not limited to the complete, endogenous amino acid sequence and may be a fragment, epitope, variant, or derivative of a protein expressed by a nucleic acid sequence.

"Amplification" refers to the production of additional copies of a sequence and is carried out using polymerase chain reaction (PCR) technologies well known in the art.

"Antibody" refers to intact molecules as well as to fragments thereof, such as Fab, $F(ab')_2$, and Fv fragments, which are capable of binding the epitopic determinant. Antibodies that bind DITHP polypeptides can be prepared using intact polypeptides or using fragments containing small peptides of interest as the immunizing antigen. The polypeptide or peptide used to immunize an animal (e.g., a mouse, a rat, or a rabbit) can be derived from the translation of RNA, or synthesized chemically, and can be conjugated to a carrier protein if desired. Commonly used carriers that are chemically coupled to peptides include bovine serum albumin, thyroglobulin, and keyhole limpet hemocyanin (KLH). The coupled peptide is then used to immunize the animal.

"Antisense sequence" refers to a sequence capable of specifically hybridizing to a target sequence. The antisense sequence may include DNA, RNA, or any nucleic acid mimic or analog such as peptide nucleic acid (PNA); oligonucleotides having modified backbone linkages such as phosphorothioates, methylphosphonates, or benzylphosphonates; oligonucleotides having modified sugar groups such as 2'-methoxyethyl sugars or 2'-methoxyethoxy sugars; or oligonucleotides having modified bases such as 5-methyl cytosine, 2'-deoxyuracil, or 7-deaza-2'-deoxyguanosine.

"Antisense sequence" refers to a sequence capable of specifically hybridizing to a target sequence. The antisense sequence can be DNA, RNA, or any nucleic acid mimic or analog.

"Antisense technology" refers to any technology which relies on the specific hybridization of an antisense sequence to a target sequence.

A "bin" is a portion of computer memory space used by a computer program for storage of data, and bounded in such a manner that data stored in a bin may be retrieved by the program.

"Biologically active" refers to an amino acid sequence having a structural, regulatory, or biochemical function of a naturally occurring amino acid sequence.

"Clone joining" is a process for combining gene bins based upon the bins' containing sequence information from the same clone. The sequences may assemble into a primary gene transcript as well as one or more splice variants.

"Complementary" describes the relationship between two single-stranded nucleic acid sequences that anneal by base-pairing (5'-A-G-T-3' pairs with its complement 3'-T-C-A-5').

A "component sequence" is a nucleic acid sequence selected by a computer program such as PHRED and used to assemble a consensus or template sequence from one or more component sequences.

A "consensus sequence" or "template sequence" is a nucleic acid sequence which has been assembled from overlapping sequences, using a computer program for fragment assembly such as the GELVIEW fragment assembly system (Genetics Computer Group (GCG), Madison WI) or using a relational database management system (RDMS).

"Conservative amino acid substitutions" are those substitutions that, when made, least interfere with the properties of the original protein, i.e., the structure and especially the function of the protein is conserved and not significantly changed by such substitutions. The table below shows amino acids which may be substituted for an original amino acid in a protein and which are regarded as conservative substitutions.

| Original Residue | Conservative Substitution |
|---|---|
| Ala | Gly, Ser |
| Arg | His, Lys |
| Asn | Asp, Gln, His |
| Asp | Asn, Glu |
| Cys | Ala, Ser |
| Gln | Asn, Glu, His |
| Glu | Asp, Gln, His |
| Gly | Ala |
| His | Asn, Arg, Gln, Glu |
| Ile | Leu, Val |
| Leu | Ile, Val |
| Lys | Arg, Gln, Glu |
| Met | Leu, Ile |
| Phe | His, Met, Leu, Trp, Tyr |
| Ser | Cys, Thr |
| Thr | Ser, Val |
| Trp | Phe, Tyr |
| Tyr | His, Phe, Trp |
| Val | Ile, Leu, Thr |

Conservative substitutions generally maintain (a) the structure of the polypeptide backbone in

the area of the substitution, for example, as a beta sheet or alpha helical conformation, (b) the charge or hydrophobicity of the molecule at the target site, or (c) the bulk of the side chain.

"Deletion" refers to a change in either a nucleic or amino acid sequence in which at least one nucleotide or amino acid residue, respectively, is absent.

"Derivative" refers to the chemical modification of a nucleic acid sequence, such as by replacement of hydrogen by an alkyl, acyl, amino, hydroxyl, or other group.

"E-value" refers to the statistical probability that a match between two sequences occurred by chance.

A "fragment" is a unique portion of dithp or DITHP which is identical in sequence to but shorter in length than the parent sequence. A fragment may comprise up to the entire length of the defined sequence, minus one nucleotide/amino acid residue. For example, a fragment may comprise from 10 to 1000 contiguous amino acid residues or nucleotides. A fragment used as a probe, primer, antigen, therapeutic molecule, or for other purposes, may be at least 5, 10, 15, 16, 20, 25, 30, 40, 50, 60, 75, 100, 150, 250 or at least 500 contiguous amino acid residues or nucleotides in length. Fragments may be preferentially selected from certain regions of a molecule. For example, a polypeptide fragment may comprise a certain length of contiguous amino acids selected from the first 250 or 500 amino acids (or first 25% or 50%) of a polypeptide as shown in a certain defined sequence. Clearly these lengths are exemplary, and any length that is supported by the specification, including the Sequence Listing and the figures, may be encompassed by the present embodiments.

A fragment of dithp comprises a region of unique polynucleotide sequence that specifically identifies dithp, for example, as distinct from any other sequence in the same genome. A fragment of dithp is useful, for example, in hybridization and amplification technologies and in analogous methods that distinguish dithp from related polynucleotide sequences. The precise length of a fragment of dithp and the region of dithp to which the fragment corresponds are routinely determinable by one of ordinary skill in the art based on the intended purpose for the fragment.

A fragment of DITHP is encoded by a fragment of dithp. A fragment of DITHP comprises a region of unique amino acid sequence that specifically identifies DITHP. For example, a fragment of DITHP is useful as an immunogenic peptide for the development of antibodies that specifically recognize DITHP. The precise length of a fragment of DITHP and the region of DITHP to which the fragment corresponds are routinely determinable by one of ordinary skill in the art based on the intended purpose for the fragment.

A "full length" nucleotide sequence is one containing at least a start site for translation to a protein sequence, followed by an open reading frame and a stop site, and encoding a "full length" polypeptide.

"Hit" refers to a sequence whose annotation will be used to describe a given template. Criteria

for selecting the top hit are as follows: if the template has one or more exact nucleic acid matches, the top hit is the exact match with highest percent identity. If the template has no exact matches but has significant protein hits, the top hit is the protein hit with the lowest E-value. If the template has no significant protein hits, but does have significant non-exact nucleotide hits, the top hit is the nucleotide hit with the lowest E-value.

"Homology" refers to sequence similarity either between a reference nucleic acid sequence and at least a fragment of a dithp or between a reference amino acid sequence and a fragment of a DITHP.

"Hybridization" refers to the process by which a strand of nucleotides anneals with a complementary strand through base pairing. Specific hybridization is an indication that two nucleic acid sequences share a high degree of identity. Specific hybridization complexes form under defined annealing conditions, and remain hybridized after the "washing" step. The defined hybridization conditions include the annealing conditions and the washing step(s), the latter of which is particularly important in determining the stringency of the hybridization process, with more stringent conditions allowing less non-specific binding, i.e., binding between pairs of nucleic acid probes that are not perfectly matched. Permissive conditions for annealing of nucleic acid sequences are routinely determinable and may be consistent among hybridization experiments, whereas wash conditions may be varied among experiments to achieve the desired stringency.

Generally, stringency of hybridization is expressed with reference to the temperature under which the wash step is carried out. Generally, such wash temperatures are selected to be about 5°C to 20°C lower than the thermal melting point ($T_m$) for the specific sequence at a defined ionic strength and pH. The $T_m$ is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. An equation for calculating $T_m$ and conditions for nucleic acid hybridization is well known and can be found in Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, 2nd ed., vol. 1-3, Cold Spring Harbor Press, Plainview NY; specifically see volume 2, chapter 9.

High stringency conditions for hybridization between polynucleotides of the present invention include wash conditions of 68°C in the presence of about 0.2 x SSC and about 0.1% SDS, for 1 hour. Alternatively, temperatures of about 65°C, 60°C, or 55°C may be used. SSC concentration may be varied from about 0.2 to 2 x SSC, with SDS being present at about 0.1%. Typically, blocking reagents are used to block non-specific hybridization. Such blocking reagents include, for instance, denatured salmon sperm DNA at about 100-200 μg/ml. Useful variations on these conditions will be readily apparent to those skilled in the art. Hybridization, particularly under high stringency conditions, may be suggestive of evolutionary similarity between the nucleotides. Such similarity is strongly indicative of a similar role for the nucleotides and their resultant proteins.

Other parameters, such as temperature, salt concentration, and detergent concentration may be

varied to achieve the desired stringency. Denaturants, such as formamide at a concentration of about 35-50% v/v, may also be used under particular circumstances, such as RNA:DNA hybridizations. Appropriate hybridization conditions are routinely determinable by one of ordinary skill in the art.

"Immunogenic" describes the potential for a natural, recombinant, or synthetic peptide, epitope, polypeptide, or protein to induce antibody production in appropriate animals, cells, or cell lines.

"Insertion" or "addition" refers to a change in either a nucleic or amino acid sequence in which at least one nucleotide or residue, respectively, is added to the sequence.

"Labeling" refers to the covalent or noncovalent joining of a polynucleotide, polypeptide, or antibody with a reporter molecule capable of producing a detectable or measurable signal.

"Microarray" is any arrangement of nucleic acids, amino acids, antibodies, etc., on a substrate. The substrate may be a solid support such as beads, glass, paper, nitrocellulose, nylon, or an appropriate membrane.

The terms "element" and "array element" refer to a polynucleotide, polypeptide, or other chemical compound having a unique and defined position on a microarray.

"Linkers" are short stretches of nucleotide sequence which may be added to a vector or a dithp to create restriction endonuclease sites to facilitate cloning. "Polylinkers" are engineered to incorporate multiple restriction enzyme sites and to provide for the use of enzymes which leave 5' or 3' overhangs (e.g., BamHI, EcoRI, and HindIII) and those which provide blunt ends (e.g., EcoRV, SnaBI, and StuI).

"Naturally occurring" refers to an endogenous polynucleotide or polypeptide that may be isolated from viruses or prokaryotic or eukaryotic cells.

"Nucleic acid sequence" refers to the specific order of nucleotides joined by phosphodiester bonds in a linear, polymeric arrangement. Depending on the number of nucleotides, the nucleic acid sequence can be considered an oligomer, oligonucleotide, or polynucleotide. The nucleic acid can be DNA, RNA, or any nucleic acid analog, such as PNA, may be of genomic or synthetic origin, may be either double-stranded or single-stranded, and can represent either the sense or antisense (complementary) strand.

"Oligomer" refers to a nucleic acid sequence of at least about 6 nucleotides and as many as about 60 nucleotides, preferably about 15 to 40 nucleotides, and most preferably between about 20 and 30 nucleotides, that may be used in hybridization or amplification technologies. Oligomers may be used as, e.g., primers for PCR, and are usually chemically synthesized.

"Operably linked" refers to the situation in which a first nucleic acid sequence is placed in a functional relationship with the second nucleic acid sequence. For instance, a promoter is operably linked to a coding sequence if the promoter affects the transcription or expression of the coding sequence. Generally, operably linked DNA sequences may be in close proximity or contiguous and, where necessary to join two protein coding regions, in the same reading frame.

"Peptide nucleic acid" (PNA) refers to a DNA mimic in which nucleotide bases are attached to a pseudopeptide backbone to increase stability. PNAs, also designated antigene agents, can prevent gene expression by targeting complementary messenger RNA.

The phrases "percent identity" and "% identity", as applied to polynucleotide sequences, refer to the percentage of residue matches between at least two polynucleotide sequences aligned using a standardized algorithm. Such an algorithm may insert, in a standardized and reproducible way, gaps in the sequences being compared in order to optimize alignment between two sequences, and therefore achieve a more meaningful comparison of the two sequences.

Percent identity between polynucleotide sequences may be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program. This program is part of the LASERGENE software package, a suite of molecular biological analysis programs (DNASTAR, Madison WI). CLUSTAL V is described in Higgins, D.G. and Sharp, P.M. (1989) CABIOS 5:151-153 and in Higgins, D.G. et al. (1992) CABIOS 8:189-191. For pairwise alignments of polynucleotide sequences, the default parameters are set as follows: Ktuple=2, gap penalty=5, window=4, and "diagonals saved"=4. The "weighted" residue weight table is selected as the default. Percent identity is reported by CLUSTAL V as the "percent similarity" between aligned polynucleotide sequence pairs.

Alternatively, a suite of commonly used and freely available sequence comparison algorithms is provided by the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) (Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410), which is available from several sources, including the NCBI, Bethesda, MD, and on the Internet at http://www.ncbi.nlm.nih.gov/BLAST/. The BLAST software suite includes various sequence analysis programs including "blastn," that is used to determine alignment between a known polynucleotide sequence and other sequences on a variety of databases. Also available is a tool called "BLAST 2 Sequences" that is used for direct pairwise comparison of two nucleotide sequences. "BLAST 2 Sequences" can be accessed and used interactively at http://www.ncbi.nlm.nih.gov/gorf/bl2/. The "BLAST 2 Sequences" tool can be used for both blastn and blastp (discussed below). BLAST programs are commonly used with gap and other parameters set to default settings. For example, to compare two nucleotide sequences, one may use blastn with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such default parameters may be, for example:

*Matrix: BLOSUM62*

*Reward for match: 1*

*Penalty for mismatch: -2*

*Open Gap: 5 and Extension Gap: 2 penalties*

*Gap x drop-off: 50*

*Expect: 10*

*Word Size: 11*

*Filter: on*

Percent identity may be measured over the length of an entire defined sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined sequence, for instance, a fragment of at least 20, at least 30, at least 40, at least 50, at least 70, at least 100, or at least 200 contiguous nucleotides. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in figures or Sequence Listings, may be used to describe a length over which percentage identity may be measured.

Nucleic acid sequences that do not show a high degree of identity may nevertheless encode similar amino acid sequences due to the degeneracy of the genetic code. It is understood that changes in nucleic acid sequence can be made using this degeneracy to produce multiple nucleic acid sequences that all encode substantially the same protein.

The phrases "percent identity" and "% identity", as applied to polypeptide sequences, refer to the percentage of residue matches between at least two polypeptide sequences aligned using a standardized algorithm. Methods of polypeptide sequence alignment are well-known. Some alignment methods take into account conservative amino acid substitutions. Such conservative substitutions, explained in more detail above, generally preserve the hydrophobicity and acidity of the substituted residue, thus preserving the structure (and therefore function) of the folded polypeptide.

Percent identity between polypeptide sequences may be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program (described and referenced above). For pairwise alignments of polypeptide sequences using CLUSTAL V, the default parameters are set as follows: Ktuple=1, gap penalty=3, window=5, and "diagonals saved"=5. The PAM250 matrix is selected as the default residue weight table. As with polynucleotide alignments, the percent identity is reported by CLUSTAL V as the "percent similarity" between aligned polypeptide sequence pairs.

Alternatively the NCBI BLAST software suite may be used. For example, for a pairwise comparison of two polypeptide sequences, one may use the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) with blastp set at default parameters. Such default parameters may be, for example:

*Matrix: BLOSUM62*

*Open Gap: 11 and Extension Gap: 1 penalty*

*Gap x drop-off: 50*

*Expect: 10*

*Word Size: 3*

*Filter: on*

Percent identity may be measured over the length of an entire defined polypeptide sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined polypeptide sequence, for instance, a fragment of at least 15, at least 20, at least 30, at least 40, at least 50, at least 70 or at least 150 contiguous residues. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in figures or Sequence Listings, may be used to describe a length over which percentage identity may be measured.

"Post-translational modification" of a DITHP may involve lipidation, glycosylation, phosphorylation, acetylation, racemization, proteolytic cleavage, and other modifications known in the art. These processes may occur synthetically or biochemically. Biochemical modifications will vary by cell type depending on the enzymatic milieu and the DITHP.

"Probe" refers to dithp or fragments thereof, which are used to detect identical, allelic or related nucleic acid sequences. Probes are isolated oligonucleotides or polynucleotides attached to a detectable label or reporter molecule. Typical labels include radioactive isotopes, ligands, chemiluminescent agents, and enzymes. "Primers" are short nucleic acids, usually DNA oligonucleotides, which may be annealed to a target polynucleotide by complementary base-pairing. The primer may then be extended along the target DNA strand by a DNA polymerase enzyme. Primer pairs can be used for amplification (and identification) of a nucleic acid sequence, e.g., by the polymerase chain reaction (PCR).

Probes and primers as used in the present invention typically comprise at least 15 contiguous nucleotides of a known sequence. In order to enhance specificity, longer probes and primers may also be employed, such as probes and primers that comprise at least 20, 30, 40, 50, 60, 70, 80, 90, 100, or at least 150 consecutive nucleotides of the disclosed nucleic acid sequences. Probes and primers may be considerably longer than these examples, and it is understood that any length supported by the specification, including the figures and Sequence Listing, may be used.

Methods for preparing and using probes and primers are described in the references, for example Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, 2nd ed., vol. 1-3, Cold Spring Harbor Press, Plainview NY; Ausubel et al.,1987, Current Protocols in Molecular Biology, Greene Publ. Assoc. & Wiley-Intersciences, New York NY; Innis et al., 1990, PCR Protocols, A Guide to Methods and Applications, Academic Press, San Diego CA. PCR primer pairs can be derived from a known sequence, for example, by using computer programs intended for that purpose such as Primer (Version 0.5, 1991, Whitehead Institute for Biomedical Research, Cambridge MA).

Oligonucleotides for use as primers are selected using software known in the art for such purpose. For example, OLIGO 4.06 software is useful for the selection of PCR primer pairs of up to 100 nucleotides each, and for the analysis of oligonucleotides and larger polynucleotides of up to 5,000

nucleotides from an input polynucleotide sequence of up to 32 kilobases. Similar primer selection programs have incorporated additional features for expanded capabilities. For example, the PrimOU primer selection program (available to the public from the Genome Center at University of Texas South West Medical Center, Dallas TX) is capable of choosing specific primers from megabase sequences and is thus useful for designing primers on a genome-wide scope. The Primer3 primer selection program (available to the public from the Whitehead Institute/MIT Center for Genome Research, Cambridge MA) allows the user to input a "mispriming library," in which sequences to avoid as primer binding sites are user-specified. Primer3 is useful, in particular, for the selection of oligonucleotides for microarrays. (The source code for the latter two primer selection programs may also be obtained from their respective sources and modified to meet the user's specific needs.) The PrimeGen program (available to the public from the UK Human Genome Mapping Project Resource Centre, Cambridge UK) designs primers based on multiple sequence alignments, thereby allowing selection of primers that hybridize to either the most conserved or least conserved regions of aligned nucleic acid sequences. Hence, this program is useful for identification of both unique and conserved oligonucleotides and polynucleotide fragments. The oligonucleotides and polynucleotide fragments identified by any of the above selection methods are useful in hybridization technologies, for example, as PCR or sequencing primers, microarray elements, or specific probes to identify fully or partially complementary polynucleotides in a sample of nucleic acids. Methods of oligonucleotide selection are not limited to those described above.

"Purified" refers to molecules, either polynucleotides or polypeptides that are isolated or separated from their natural environment and are at least 60% free, preferably at least 75% free, and most preferably at least 90% free from other compounds with which they are naturally associated.

A "recombinant nucleic acid" is a sequence that is not naturally occurring or has a sequence that is made by an artificial combination of two or more otherwise separated segments of sequence. This artificial combination is often accomplished by chemical synthesis or, more commonly, by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques such as those described in Sambrook, supra. The term recombinant includes nucleic acids that have been altered solely by addition, substitution, or deletion of a portion of the nucleic acid. Frequently, a recombinant nucleic acid may include a nucleic acid sequence operably linked to a promoter sequence. Such a recombinant nucleic acid may be part of a vector that is used, for example, to transform a cell.

Alternatively, such recombinant nucleic acids may be part of a viral vector, e.g., based on a vaccinia virus, that could be use to vaccinate a mammal wherein the recombinant nucleic acid is expressed, inducing a protective immunological response in the mammal.

"Regulatory element" refers to a nucleic acid sequence from nontranslated regions of a gene, and includes enhancers, promoters, introns, and 3' untranslated regions, which interact with host

proteins to carry out or regulate transcription or translation.

"Reporter" molecules are chemical or biochemical moieties used for labeling a nucleic acid, an amino acid, or an antibody. They include radionuclides; enzymes; fluorescent, chemiluminescent, or chromogenic agents; substrates; cofactors; inhibitors; magnetic particles; and other moieties known in the art.

An "RNA equivalent," in reference to a DNA sequence, is composed of the same linear sequence of nucleotides as the reference DNA sequence with the exception that all occurrences of the nitrogenous base thymine are replaced with uracil, and the sugar backbone is composed of ribose instead of deoxyribose.

"Sample" is used in its broadest sense. Samples may contain nucleic or amino acids, antibodies, or other materials, and may be derived from any source (e.g., bodily fluids including, but not limited to, saliva, blood, and urine; chromosome(s), organelles, or membranes isolated from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; and cleared cells or tissues or blots or imprints from such cells or tissues).

"Specific binding" or "specifically binding" refers to the interaction between a protein or peptide and its agonist, antibody, antagonist, or other binding partner. The interaction is dependent upon the presence of a particular structure of the protein, e.g., the antigenic determinant or epitope, recognized by the binding molecule. For example, if an antibody is specific for epitope "A," the presence of a polypeptide containing epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

"Substitution" refers to the replacement of at least one nucleotide or amino acid by a different nucleotide or amino acid.

"Substrate" refers to any suitable rigid or semi-rigid support including, e.g., membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, tubing, plates, polymers, microparticles or capillaries. The substrate can have a variety of surface forms, such as wells, trenches, pins, channels and pores, to which polynucleotides or polypeptides are bound.

A "transcript image" refers to the collective pattern of gene expression by a particular tissue or cell type under given conditions at a given time.

"Transformation" refers to a process by which exogenous DNA enters a recipient cell. Transformation may occur under natural or artificial conditions using various methods well known in the art. Transformation may rely on any known method for the insertion of foreign nucleic acid sequences into a prokaryotic or eukaryotic host cell. The method is selected based on the host cell being transformed.

"Transformants" include stably transformed cells in which the inserted DNA is capable of

replication either as an autonomously replicating plasmid or as part of the host chromosome, as well as cells which transiently express inserted DNA or RNA.

A "transgenic organism," as used herein, is any organism, including but not limited to animals and plants, in which one or more of the cells of the organism contains heterologous nucleic acid introduced by way of human intervention, such as by transgenic techniques well known in the art. The nucleic acid is introduced into the cell, directly or indirectly by introduction into a precursor of the cell, by way of deliberate genetic manipulation, such as by microinjection or by infection with a recombinant virus. The term genetic manipulation does not include classical cross-breeding, or in vitro fertilization, but rather is directed to the introduction of a recombinant DNA molecule. The transgenic organisms contemplated in accordance with the present invention include bacteria, cyanobacteria, fungi, and plants and animals. The isolated DNA of the present invention can be introduced into the host by methods known in the art, for example infection, transfection, transformation or transconjugation. Techniques for transferring the DNA of the present invention into such organisms are widely known and provided in references such as Sambrook et al. (1989), supra.

A "variant" of a particular nucleic acid sequence is defined as a nucleic acid sequence having at least 25% sequence identity to the particular nucleic acid sequence over a certain length of one of the nucleic acid sequences using blastn with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of nucleic acids may show, for example, at least 30%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95% or even at least 98% or greater sequence identity over a certain defined length. The variant may result in "conservative" amino acid changes which do not affect structural and/or chemical properties. A variant may be described as, for example, an "allelic" (as defined above), "splice," "species," or "polymorphic" variant. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing of exons during mRNA processing. The corresponding polypeptide may possess additional functional domains or lack domains that are present in the reference molecule. Species variants are polynucleotide sequences that vary from one species to another. The resulting polypeptides generally will have significant amino acid identity relative to each other. A polymorphic variant is a variation in the polynucleotide sequence of a particular gene between individuals of a given species. Polymorphic variants also may encompass "single nucleotide polymorphisms" (SNPs) in which the polynucleotide sequence varies by one base. The presence of SNPs may be indicative of, for example, a certain population, a disease state, or a propensity for a disease state.

In an alternative, variants of the polynucleotides of the present invention may be generated through recombinant methods. One possible method is a DNA shuffling technique such as MOLECULARBREEDING (Maxygen Inc., Santa Clara CA; described in U.S. Patent Number

5,837,458; Chang, C.-C. et al. (1999) Nat. Biotechnol. 17:793-797; Christians, F.C. et al. (1999) Nat. Biotechnol. 17:259-264; and Crameri, A. et al. (1996) Nat. Biotechnol. 14:315-319) to alter or improve the biological properties of DITHP, such as its biological or enzymatic activity or its ability to bind to other molecules or compounds. DNA shuffling is a process by which a library of gene variants is produced using PCR-mediated recombination of gene fragments. The library is then subjected to selection or screening procedures that identify those gene variants with the desired properties. These preferred variants may then be pooled and further subjected to recursive rounds of DNA shuffling and selection/screening. Thus, genetic diversity is created through "artificial" breeding and rapid molecular evolution. For example, fragments of a single gene containing random point mutations may be recombined, screened, and then reshuffled until the desired properties are optimized. Alternatively, fragments of a given gene may be recombined with fragments of homologous genes in the same gene family, either from the same or different species, thereby maximizing the genetic diversity of multiple naturally occurring genes in a directed and controllable manner.

A "variant" of a particular polypeptide sequence is defined as a polypeptide sequence having at least 40% sequence identity to the particular polypeptide sequence over a certain length of one of the polypeptide sequences using blastp with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of polypeptides may show, for example, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, or at least 98% or greater sequence identity over a certain defined length of one of the polypeptides.

## THE INVENTION

In a particular embodiment, cDNA sequences derived from human tissues and cell lines were aligned based on nucleotide sequence identity and assembled into "consensus" or "template" sequences which are designated by the template identification numbers (template IDs) in column 2 of Table 1. The sequence identification numbers (SEQ ID NO:s) corresponding to the template IDs are shown in column 1. The template sequences have similarity to GenBank sequences, or "hits," as designated by the GI Numbers in column 3. The statistical probability of each GenBank hit is indicated by a probability score in column 4, and the functional annotation corresponding to each GenBank hit is listed in column 5.

The invention incorporates the nucleic acid sequences of these templates as disclosed in the Sequence Listing and the use of these sequences in the diagnosis and treatment of disease states characterized by defects in human molecules. The invention further utilizes these sequences in hybridization and amplification technologies, and in particular, in technologies which assess gene expression patterns correlated with specific cells or tissues and their responses in vivo or in vitro to pharmaceutical agents, toxins, and other treatments. In this manner, the sequences of the present

invention are used to develop a transcript image for a particular cell or tissue.

Derivation of Nucleic Acid Sequences

cDNA was isolated from libraries constructed using RNA derived from normal and diseased human tissues and cell lines. The human tissues and cell lines used for cDNA library construction were selected from a broad range of sources to provide a diverse population of cDNAs representative of gene transcription throughout the human body. Descriptions of the human tissues and cell lines used for cDNA library construction are provided in the LIFESEQ database (Incyte Genomics, Inc. (Incyte), Palo Alto CA). Human tissues were broadly selected from, for example, cardiovascular, dermatologic, endocrine, gastrointestinal, hematopoietic/immune system, musculoskeletal, neural, reproductive, and urologic sources.

Cell lines used for cDNA library construction were derived from, for example, leukemic cells, teratocarcinomas, neuroepitheliomas, cervical carcinoma, lung fibroblasts, and endothelial cells. Such cell lines include, for example, THP-1, Jurkat, HUVEC, hNT2, WI38, HeLa, and other cell lines commonly used and available from public depositories (American Type Culture Collection, Manassas VA). Prior to mRNA isolation, cell lines were untreated, treated with a pharmaceutical agent such as 5'-aza-2'-deoxycytidine, treated with an activating agent such as lipopolysaccharide in the case of leukocytic cell lines, or, in the case of endothelial cell lines, subjected to shear stress.

Sequencing of the cDNAs

Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ the Klenow fragment of DNA polymerase I, SEQUENASE DNA polymerase (U.S. Biochemical Corporation, Cleveland OH), Taq polymerase (PE Biosystems, Foster City CA), thermostable T7 polymerase (Amersham Pharmacia Biotech, Inc. (Amersham Pharmacia Biotech), Piscataway NJ), or combinations of polymerases and proofreading exonucleases such as those found in the ELONGASE amplification system (Life Technologies Inc. (Life Technologies), Gaithersburg MD), to extend the nucleic acid sequence from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single-stranded and double-stranded templates. Chain termination reaction products may be electrophoresed on urea-polyacrylamide gels and detected either by autoradiography (for radioisotope-labeled nucleotides) or by fluorescence (for fluorophore-labeled nucleotides). Automated methods for mechanized reaction preparation, sequencing, and analysis using fluorescence detection methods have been developed. Machines used to prepare cDNAs for sequencing can include the MICROLAB 2200 liquid transfer system (Hamilton Company (Hamilton), Reno NV), Peltier thermal cycler (PTC200; MJ Research, Inc. (MJ Research), Watertown MA), and ABI CATALYST 800 thermal cycler (PE Biosystems). Sequencing can be carried out using,

for example, the ABI 373 or 377 (PE Biosystems) or MEGABACE 1000 (Molecular Dynamics, Inc. (Molecular Dynamics), Sunnyvale CA) DNA sequencing systems, or other automated and manual sequencing systems well known in the art.

The nucleotide sequences of the Sequence Listing have been prepared by current, state-of-the-art, automated methods and, as such, may contain occasional sequencing errors or unidentified nucleotides. Such unidentified nucleotides are designated by an N. These infrequent unidentified bases do not represent a hindrance to practicing the invention for those skilled in the art. Several methods employing standard recombinant techniques may be used to correct errors and complete the missing sequence information. (See, e.g., those described in Ausubel, F.M. et al. (1997) Short Protocols in Molecular Biology, John Wiley & Sons, New York NY; and Sambrook, J. et al. (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY.)

Assembly of cDNA Sequences

Human polynucleotide sequences may be assembled using programs or algorithms well known in the art. Sequences to be assembled are related, wholly or in part, and may be derived from a single or many different transcripts. Assembly of the sequences can be performed using such programs as PHRAP (Phils Revised Assembly Program) and the GELVIEW fragment assembly system (GCG), or other methods known in the art.

Alternatively, cDNA sequences are used as "component" sequences that are assembled into "template" or "consensus" sequences as follows. Sequence chromatograms are processed, verified, and quality scores are obtained using PHRED. Raw sequences are edited using an editing pathway known as Block 1 (See, e.g., the LIFESEQ Assembled User Guide, Incyte Genomics, Palo Alto, CA). A series of BLAST comparisons is performed and low-information segments and repetitive elements (e.g., dinucleotide repeats, Alu repeats, etc.) are replaced by "n's", or masked, to prevent spurious matches. Mitochondrial and ribosomal RNA sequences are also removed. The processed sequences are then loaded into a relational database management system (RDMS) which assigns edited sequences to existing templates, if available. When additional sequences are added into the RDMS, a process is initiated which modifies existing templates or creates new templates from works in progress (i.e., nonfinal assembled sequences) containing queued sequences or the sequences themselves. After the new sequences have been assigned to templates, the templates can be merged into bins. If multiple templates exist in one bin, the bin can be split and the templates reannotated.

Once gene bins have been generated based upon sequence alignments, bins are "clone joined" based upon clone information. Clone joining occurs when the 5' sequence of one clone is present in one bin and the 3' sequence from the same clone is present in a different bin, indicating that the two bins should be merged into a single bin. Only bins which share at least two different clones are merged.

A resultant template sequence may contain either a partial or a full length open reading frame, or all or part of a genetic regulatory element. This variation is due in part to the fact that the full length cDNAs of many genes are several hundred, and sometimes several thousand, bases in length. With current technology, cDNAs comprising the coding regions of large genes cannot be cloned because of vector limitations, incomplete reverse transcription of the mRNA, or incomplete "second strand" synthesis. Template sequences may be extended to include additional contiguous sequences derived from the parent RNA transcript using a variety of methods known to those of skill in the art. Extension may thus be used to achieve the full length coding sequence of a gene.

Analysis of the cDNA Sequences

The cDNA sequences are analyzed using a variety of programs and algorithms which are well known in the art. (See, e.g., Ausubel, 1997, supra, Chapter 7.7; Meyers, R.A. (Ed.) (1995) Molecular Biology and Biotechnology, Wiley VCH, New York NY, pp. 856-853; and Table 5.) These analyses comprise both reading frame determinations, e.g., based on triplet codon periodicity for particular organisms (Fickett, J.W. (1982) Nucleic Acids Res. 10:5303-5318); analyses of potential start and stop codons; and homology searches.

Computer programs known to those of skill in the art for performing computer-assisted searches for amino acid and nucleic acid sequence similarity, include, for example, Basic Local Alignment Search Tool (BLAST; Altschul, S.F. (1993) J. Mol. Evol. 36:290-300; Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410). BLAST is especially useful in determining exact matches and comparing two sequence fragments of arbitrary but equal lengths, whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score set by the user (Karlin, S. et al. (1988) Proc. Natl. Acad. Sci. USA 85:841-845). Using an appropriate search tool (e.g., BLAST or HMM), GenBank, SwissProt, BLOCKS, PFAM and other databases may be searched for sequences containing regions of homology to a query dithp or DITHP of the present invention.

Other approaches to the identification, assembly, storage, and display of nucleotide and polypeptide sequences are provided in "Relational Database for Storing Biomolecule Information," U.S.S.N. 08/947,845, filed October 9, 1997; "Project-Based Full-Length Biomolecular Sequence Database," U.S.S.N. 08/811,758, filed March 6, 1997; and "Relational Database and System for Storing Information Relating to Biomolecular Sequences," U.S.S.N. 09/034,807, filed March 4, 1998, all of which are incorporated by reference herein in their entirety.

Protein hierarchies can be assigned to the putative encoded polypeptide based on, e.g., motif, BLAST, or biological analysis. Methods for assigning these hierarchies are described, for example, in "Database System Employing Protein Function Hierarchies for Viewing Biomolecular Sequence Data," U.S.S.N. 08/812,290, filed March 6, 1997, incorporated herein by reference.

Identification of Human Diagnostic and Therapeutic Molecules Encoded by dithp

The identities of the DITHP encoded by the dithp of the present invention were obtained by analysis of the assembled cDNA sequences. SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, and SEQ ID NO:5 encode, for example, human enzyme molecules. SEQ ID NO:6 and SEQ ID NO:7 encode, for example, receptor molecules. SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, and SEQ ID NO:12 encode, for example, intracellular signaling molecules. SEQ ID NO:13 encodes, for example, a membrane transport molecule. SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, and SEQ ID NO:20 encode, for example, nucleic acid synthesis and modification molecules. SEQ ID NO:21 and SEQ ID NO:22 encode, for example, adhesion molecules. SEQ ID NO:23 and SEQ ID NO:24 encode, for example, electron transfer associated molecules. SEQ ID NO:25 encodes, for example, a secreted/extracellular matrix molecule. SEQ ID NO:26 and SEQ ID NO:27 encode, for example, cytoskeletal molecules. SEQ ID NO:28 and SEQ ID NO:29 encode, for example, cell membrane molecules. SEQ ID NO:30 and SEQ ID NO:31 encode, for example, ribosomal molecules. SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, SEQ ID NO:35, SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, and SEQ ID NO:43 encode, for example, transcription factor molecules. SEQ ID NO:44, SEQ ID NO:45, and SEQ ID NO:46 encode, for example, organelle associated molecules. SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, and SEQ ID NO:50 encode, for example, biochemical pathway molecules. SEQ ID NO:51 and SEQ ID NO:52 encode, for example, molecules associated with growth and development.

Sequences of Human Diagnostic and Therapeutic Molecules

The dithp of the present invention may be used for a variety of diagnostic and therapeutic purposes. For example, a dithp may be used to diagnose a particular condition, disease, or disorder associated with human molecules. Such conditions, diseases, and disorders include, but are not limited to, a cell proliferative disorder, such as actinic keratosis, arteriosclerosis, atherosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, a cancer of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus; an autoimmune/inflammatory disorder, such as inflammation, actinic keratosis, acquired immunodeficiency syndrome (AIDS), Addison's disease, adult respiratory distress syndrome,

allergies, ankylosing spondylitis, amyloidosis, anemia, arteriosclerosis, asthma, atherosclerosis, autoimmune hemolytic anemia, autoimmune thyroiditis, bronchitis, bursitis, cholecystitis, cirrhosis, contact dermatitis, Crohn's disease, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, erythroblastosis fetalis, erythema nodosum, atrophic gastritis, glomerulonephritis, Goodpasture's syndrome, gout, Graves' disease, Hashimoto's thyroiditis, paroxysmal nocturnal hemoglobinuria, hepatitis, hypereosinophilia, irritable bowel syndrome, episodic lymphopenia with lymphocytotoxins, mixed connective tissue disease (MCTD), multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, myelofibrosis, osteoarthritis, osteoporosis, pancreatitis, polycythemia vera, polymyositis, psoriasis, Reiter's syndrome, rheumatoid arthritis, scleroderma, Sjögren's syndrome, systemic anaphylaxis, systemic lupus erythematosus, systemic sclerosis, primary thrombocythemia, thrombocytopenic purpura, ulcerative colitis, uveitis, Werner syndrome, complications of cancer, hemodialysis, and extracorporeal circulation, trauma, and hematopoietic cancer including lymphoma, leukemia, and myeloma; an infection caused by a viral agent classified as adenovirus, arenavirus, bunyavirus, calicivirus, coronavirus, filovirus, hepadnavirus, herpesvirus, flavivirus, orthomyxovirus, parvovirus, papovavirus, paramyxovirus, picornavirus, poxvirus, reovirus, retrovirus, rhabdovirus, or togavirus; an infection caused by a bacterial agent classified as pneumococcus, staphylococcus, streptococcus, bacillus, corynebacterium, clostridium, meningococcus, gonococcus, listeria, moraxella, kingella, haemophilus, legionella, bordetella, gram-negative enterobacterium including shigella, salmonella, or campylobacter, pseudomonas, vibrio, brucella, francisella, yersinia, bartonella, norcardium, actinomyces, mycobacterium, spirochaetale, rickettsia, chlamydia, or mycoplasma; an infection caused by a fungal agent classified as aspergillus, blastomyces, dermatophytes, cryptococcus, coccidioides, malasezzia, histoplasma, or other mycosis-causing fungal agent; and an infection caused by a parasite classified as plasmodium or malaria-causing, parasitic entamoeba, leishmania, trypanosoma, toxoplasma, pneumocystis carinii, intestinal protozoa such as giardia, trichomonas, tissue nematode such as trichinella, intestinal nematode such as ascaris, lymphatic filarial nematode, trematode such as schistosoma, and cestrode such as tapeworm; a developmental disorder such as renal tubular acidosis, anemia, Cushing's syndrome, achondroplastic dwarfism, Duchenne and Becker muscular dystrophy, epilepsy, gonadal dysgenesis, WAGR syndrome (Wilms' tumor, aniridia, genitourinary abnormalities, and mental retardation), Smith-Magenis syndrome, myelodysplastic syndrome, hereditary mucoepithelial dysplasia, hereditary keratodermas, hereditary neuropathies such as Charcot-Marie-Tooth disease and neurofibromatosis, hypothyroidism, hydrocephalus, seizure disorders such as Syndenham's chorea and cerebral palsy, spina bifida, anencephaly, craniorachischisis, congenital glaucoma, cataract, and sensorineural hearing loss; an endocrine disorder such as a disorder of the hypothalamus and/or pituitary resulting from lesions such as a primary brain tumor, adenoma, infarction associated with pregnancy,

hypophysectomy, aneurysm, vascular malformation, thrombosis, infection, immunological disorder, and complication due to head trauma; a disorder associated with hypopituitarism including hypogonadism, Sheehan syndrome, diabetes insipidus, Kallman's disease, Hand-Schuller-Christian disease, Letterer-Siwe disease, sarcoidosis, empty sella syndrome, and dwarfism; a disorder associated with hyperpituitarism including acromegaly, giantism, and syndrome of inappropriate antidiuretic hormone (ADH) secretion (SIADH) often caused by benign adenoma; a disorder associated with hypothyroidism including goiter, myxedema, acute thyroiditis associated with bacterial infection, subacute thyroiditis associated with viral infection, autoimmune thyroiditis (Hashimoto's disease), and cretinism; a disorder associated with hyperthyroidism including thyrotoxicosis and its various forms, Grave's disease, pretibial myxedema, toxic multinodular goiter, thyroid carcinoma, and Plummer's disease; a disorder associated with hyperparathyroidism including Conn disease (chronic hypercalemia); a pancreatic disorder such as Type I or Type II diabetes mellitus and associated complications; a disorder associated with the adrenals such as hyperplasia, carcinoma, or adenoma of the adrenal cortex, hypertension associated with alkalosis, amyloidosis, hypokalemia, Cushing's disease, Liddle's syndrome, and Arnold-Healy-Gordon syndrome, pheochromocytoma tumors, and Addison's disease; a disorder associated with gonadal steroid hormones such as: in women, abnormal prolactin production, infertility, endometriosis, perturbation of the menstrual cycle, polycystic ovarian disease, hyperprolactinemia, isolated gonadotropin deficiency, amenorrhea, galactorrhea, hermaphroditism, hirsutism and virilization, breast cancer, and, in post-menopausal women, osteoporosis; and, in men, Leydig cell deficiency, male climacteric phase, and germinal cell aplasia, a hypergonadal disorder associated with Leydig cell tumors, androgen resistance associated with absence of androgen receptors, syndrome of 5 α-reductase, and gynecomastia; a metabolic disorder such as Addison's disease, cerebrotendinous xanthomatosis, congenital adrenal hyperplasia, coumarin resistance, cystic fibrosis, diabetes, fatty hepatocirrhosis, fructose-1,6-diphosphatase deficiency, galactosemia, goiter, glucagonoma, glycogen storage diseases, hereditary fructose intolerance, hyperadrenalism, hypoadrenalism, hyperparathyroidism, hypoparathyroidism, hypercholesterolemia, hyperthyroidism, hypoglycemia, hypothyroidism, hyperlipidemia, hyperlipemia, lipid myopathies, lipodystrophies, lysosomal storage diseases, mannosidosis, neuraminidase deficiency, obesity, pentosuria phenylketonuria, pseudovitamin D-deficiency rickets; disorders of carbohydrate metabolism such as congenital type II dyserythropoietic anemia, diabetes, insulin-dependent diabetes mellitus, non-insulin-dependent diabetes mellitus, fructose-1,6-diphosphatase deficiency, galactosemia, glucagonoma, hereditary fructose intolerance, hypoglycemia, mannosidosis, neuraminidase deficiency, obesity, galactose epimerase deficiency, glycogen storage diseases, lysosomal storage diseases, fructosuria, pentosuria, and inherited abnormalities of pyruvate metabolism; disorders of lipid metabolism such as fatty liver, cholestasis, primary biliary cirrhosis, carnitine deficiency,

carnitine palmitoyltransferase deficiency, myoadenylate deaminase deficiency, hypertriglyceridemia, lipid storage disorders such Fabry's disease, Gaucher's disease, Niemann-Pick's disease, metachromatic leukodystrophy, adrenoleukodystrophy, $GM_2$ gangliosidosis, and ceroid lipofuscinosis, abetalipoproteinemia, Tangier disease, hyperlipoproteinemia, diabetes mellitus, lipodystrophy, lipomatoses, acute panniculitis, disseminated fat necrosis, adiposis dolorosa, lipoid adrenal hyperplasia, minimal change disease, lipomas, atherosclerosis, hypercholesterolemia, hypercholesterolemia with hypertriglyceridemia, primary hypoalphalipoproteinemia, hypothyroidism, renal disease, liver disease, lecithin:cholesterol acyltransferase deficiency, cerebrotendinous xanthomatosis, sitosterolemia, hypocholesterolemia, Tay-Sachs disease, Sandhoff's disease, hyperlipidemia, hyperlipemia, lipid myopathies, and obesity; and disorders of copper metabolism such as Menke's disease, Wilson's disease, and Ehlers-Danlos syndrome type IX; a neurological disorder such as epilepsy, ischemic cerebrovascular disease, stroke, cerebral neoplasms, Alzheimer's disease, Pick's disease, Huntington's disease, dementia, Parkinson's disease and other extrapyramidal disorders, amyotrophic lateral sclerosis and other motor neuron disorders, progressive neural muscular atrophy, retinitis pigmentosa, hereditary ataxias, multiple sclerosis and other demyelinating diseases, bacterial and viral meningitis, brain abscess, subdural empyema, epidural abscess, suppurative intracranial thrombophlebitis, myelitis and radiculitis, viral central nervous system disease, prion diseases including kuru, Creutzfeldt-Jakob disease, and Gerstmann-Straussler-Scheinker syndrome, fatal familial insomnia, nutritional and metabolic diseases of the nervous system, neurofibromatosis, tuberous sclerosis, cerebelloretinal hemangioblastomatosis, encephalotrigeminal syndrome, mental retardation and other developmental disorder of the central nervous system, cerebral palsy, a neuroskeletal disorder, an autonomic nervous system disorder, a cranial nerve disorder, a spinal cord disease, muscular dystrophy and other neuromuscular disorder, a peripheral nervous system disorder, dermatomyositis and polymyositis, inherited, metabolic, endocrine, and toxic myopathy, myasthenia gravis, periodic paralysis, a mental disorder including mood, anxiety, and schizophrenic disorders, seasonal affective disorder (SAD), akathesia, amnesia, catatonia, diabetic neuropathy, tardive dyskinesia, dystonias, paranoid psychoses, postherpetic neuralgia, and Tourette's disorder; a gastrointestinal disorder including ulcerative colitis, gastric and duodenal ulcers, cystinuria, dibasicaminoaciduria, hypercystinuria, lysinuria, hartnup disease, tryptophan malabsorption, methionine malabsorption, histidinuria, iminoglycinuria, dicarboxylicaminoaciduria, cystinosis, renal glycosuria, hypouricemia, familial hypophophatemic rickets, congenital chloridorrhea, distal renal tubular acidosis, Menkes' disease, Wilson's disease, lethal diarrhea, juvenile pernicious anemia, folate malabsorption, adrenoleukodystrophy, hereditary myoglobinuria, and Zellweger syndrome; a transport disorder such as akinesia, amyotrophic lateral sclerosis, ataxia telangiectasia, cystic fibrosis, Becker's muscular dystrophy, Bell's palsy, Charcot-Marie Tooth disease, diabetes mellitus, diabetes

insipidus, diabetic neuropathy, Duchenne muscular dystrophy, hyperkalemic periodic paralysis, normokalemic periodic paralysis, Parkinson's disease, malignant hyperthermia, multidrug resistance, myasthenia gravis, myotonic dystrophy, catatonia, tardive dyskinesia, dystonias, peripheral neuropathy, cerebral neoplasms, prostate cancer, cardiac disorders associated with transport, e.g., angina, bradyarrythmia, tachyarrythmia, hypertension, Long QT syndrome, myocarditis, cardiomyopathy, nemaline myopathy, centronuclear myopathy, lipid myopathy, mitochondrial myopathy, thyrotoxic myopathy, ethanol myopathy, dermatomyositis, inclusion body myositis, infectious myositis, and polymyositis, neurological disorders associated with transport, e.g., Alzheimer's disease, amnesia, bipolar disorder, dementia, depression, epilepsy, Tourette's disorder, paranoid psychoses, and schizophrenia, and other disorders associated with transport, e.g., neurofibromatosis, postherpetic neuralgia, trigeminal neuropathy, sarcoidosis, sickle cell anemia, cataracts, infertility, pulmonary artery stenosis, sensorineural autosomal deafness, hyperglycemia, hypoglycemia, Grave's disease, goiter, glucose-galactose malabsorption syndrome, hypercholesterolemia, Cushing's disease, and Addison's disease; and a connective tissue disorder such as osteogenesis imperfecta, Ehlers-Danlos syndrome, chondrodysplasias, Marfan syndrome, Alport syndrome, familial aortic aneurysm, achondroplasia, mucopolysaccharidoses, osteoporosis, osteopetrosis, Paget's disease, rickets, osteomalacia, hyperparathyroidism, renal osteodystrophy, osteonecrosis, osteomyelitis, osteoma, osteoid osteoma, osteoblastoma, osteosarcoma, osteochondroma, chondroma, chondroblastoma, chondromyxoid fibroma, chondrosarcoma, fibrous cortical defect, nonossifying fibroma, fibrous dysplasia, fibrosarcoma, malignant fibrous histiocytoma, Ewing's sarcoma, primitive neuroectodermal tumor, giant cell tumor, osteoarthritis, rheumatoid arthritis, ankylosing spondyloarthritis, Reiter's syndrome, psoriatic arthritis, enteropathic arthritis, infectious arthritis, gout, gouty arthritis, calcium pyrophosphate crystal deposition disease, ganglion, synovial cyst, villonodular synovitis, systemic sclerosis, Dupuytren's contracture, hepatic fibrosis, lupus erythematosus, mixed connective tissue disease, epidermolysis bullosa simplex, bullous congenital ichthyosiform erythroderma (epidermolytic hyperkeratosis), non-epidermolytic and epidermolytic palmoplantar keratoderma, ichthyosis bullosa of Siemens, pachyonychia congenita, and white sponge nevus. The dithp can be used to detect the presence of, or to quantify the amount of, a dithp-related polynucleotide in a sample. This information is then compared to information obtained from appropriate reference samples, and a diagnosis is established. Alternatively, a polynucleotide complementary to a given dithp can inhibit or inactivate a therapeutically relevant gene related to the dithp.

Analysis of dithp Expression Patterns

The expression of dithp may be routinely assessed by hybridization-based methods to

determine, for example, the tissue-specificity, disease-specificity, or developmental stage-specificity of dithp expression. For example, the level of expression of dithp may be compared among different cell types or tissues, among diseased and normal cell types or tissues, among cell types or tissues at different developmental stages, or among cell types or tissues undergoing various treatments. This type of analysis is useful, for example, to assess the relative levels of dithp expression in fully or partially differentiated cells or tissues, to determine if changes in dithp expression levels are correlated with the development or progression of specific disease states, and to assess the response of a cell or tissue to a specific therapy, for example, in pharmacological or toxicological studies. Methods for the analysis of dithp expression are based on hybridization and amplification technologies and include membrane-based procedures such as northern blot analysis, high-throughput procedures that utilize, for example, microarrays, and PCR-based procedures.

Hybridization and Genetic Analysis

The dithp, their fragments, or complementary sequences, may be used to identify the presence of and/or to determine the degree of similarity between two (or more) nucleic acid sequences. The dithp may be hybridized to naturally occurring or recombinant nucleic acid sequences under appropriately selected temperatures and salt concentrations. Hybridization with a probe based on the nucleic acid sequence of at least one of the dithp allows for the detection of nucleic acid sequences, including genomic sequences, which are identical or related to the dithp of the Sequence Listing. Probes may be selected from non-conserved or unique regions of at least one of the polynucleotides of SEQ ID NO:1-52 and tested for their ability to identify or amplify the target nucleic acid sequence using standard protocols.

Polynucleotide sequences that are capable of hybridizing, in particular, to those shown in SEQ ID NO:1-52 and fragments thereof, can be identified using various conditions of stringency. (See, e.g., Wahl, G.M. and S.L. Berger (1987) Methods Enzymol. 152:399-407; Kimmel, A.R. (1987) Methods Enzymol. 152:507-511.) Hybridization conditions are discussed in "Definitions."

A probe for use in Southern or northern hybridization may be derived from a fragment of a dithp sequence, or its complement, that is up to several hundred nucleotides in length and is either single-stranded or double-stranded. Such probes may be hybridized in solution to biological materials such as plasmids, bacterial, yeast, or human artificial chromosomes, cleared or sectioned tissues, or to artificial substrates containing dithp. Microarrays are particularly suitable for identifying the presence of and detecting the level of expression for multiple genes of interest by examining gene expression correlated with, e.g., various stages of development, treatment with a drug or compound, or disease progression. An array analogous to a dot or slot blot may be used to arrange and link polynucleotides to the surface of a substrate using one or more of the following: mechanical (vacuum), chemical,

thermal, or UV bonding procedures. Such an array may contain any number of dithp and may be produced by hand or by using available devices, materials, and machines.

Microarrays may be prepared, used, and analyzed using methods known in the art. (See, e.g., Brennan, T.M. et al. (1995) U.S. Patent No. 5,474,796; Schena, M. et al. (1996) Proc. Natl. Acad. Sci. USA 93:10614-10619; Baldeschweiler et al. (1995) PCT application WO95/251116; Shalon, D. et al. (1995) PCT application WO95/35505; Heller, R.A. et al. (1997) Proc. Natl. Acad. Sci. USA 94:2150-2155; and Heller, M.J. et al. (1997) U.S. Patent No. 5,605,662.)

Probes may be labeled by either PCR or enzymatic techniques using a variety of commercially available reporter molecules. For example, commercial kits are available for radioactive and chemiluminescent labeling (Amersham Pharmacia Biotech) and for alkaline phosphatase labeling (Life Technologies). Alternatively, dithp may be cloned into commercially available vectors for the production of RNA probes. Such probes may be transcribed in the presence of at least one labeled nucleotide (e.g., $^{32}$P-ATP, Amersham Pharmacia Biotech).

Additionally the polynucleotides of SEQ ID NO:1-52 or suitable fragments thereof can be used to isolate full length cDNA sequences utilizing hybridization and/or amplification procedures well known in the art, e.g., cDNA library screening, PCR amplification, etc. The molecular cloning of such full length cDNA sequences may employ the method of cDNA library screening with probes using the hybridization, stringency, washing, and probing strategies described above and in Ausubel, supra, Chapters 3, 5, and 6. These procedures may also be employed with genomic libraries to isolate genomic sequences of dithp in order to analyze, e.g., regulatory elements.

Genetic Mapping

Gene identification and mapping are important in the investigation and treatment of almost all conditions, diseases, and disorders. Cancer, cardiovascular disease, Alzheimer's disease, arthritis, diabetes, and mental illnesses are of particular interest. Each of these conditions is more complex than the single gene defects of sickle cell anemia or cystic fibrosis, with select groups of genes being predictive of predisposition for a particular condition, disease, or disorder. For example, cardiovascular disease may result from malfunctioning receptor molecules that fail to clear cholesterol from the bloodstream, and diabetes may result when a particular individual's immune system is activated by an infection and attacks the insulin-producing cells of the pancreas. In some studies, Alzheimer's disease has been linked to a gene on chromosome 21; other studies predict a different gene and location. Mapping of disease genes is a complex and reiterative process and generally proceeds from genetic linkage analysis to physical mapping.

As a condition is noted among members of a family, a genetic linkage map traces parts of chromosomes that are inherited in the same pattern as the condition. Statistics link the inheritance of

particular conditions to particular regions of chromosomes, as defined by RFLP or other markers. (See, for example, Lander, E. S. and Botstein, D. (1986) Proc. Natl. Acad. Sci. USA 83:7353-7357.) Occasionally, genetic markers and their locations are known from previous studies. More often, however, the markers are simply stretches of DNA that differ among individuals. Examples of genetic linkage maps can be found in various scientific journals or at the Online Mendelian Inheritance in Man (OMIM) World Wide Web site.

In another embodiment of the invention, dithp sequences may be used to generate hybridization probes useful in chromosomal mapping of naturally occurring genomic sequences. Either coding or noncoding sequences of dithp may be used, and in some instances, noncoding sequences may be preferable over coding sequences. For example, conservation of a dithp coding sequence among members of a multi-gene family may potentially cause undesired cross hybridization during chromosomal mapping. The sequences may be mapped to a particular chromosome, to a specific region of a chromosome, or to artificial chromosome constructions, e.g., human artificial chromosomes (HACs), yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), bacterial P1 constructions, or single chromosome cDNA libraries. (See, e.g., Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355; Price, C.M. (1993) Blood Rev. 7:127-134; and Trask, B.J. (1991) Trends Genet. 7:149-154.)

Fluorescent in situ hybridization (FISH) may be correlated with other physical chromosome mapping techniques and genetic map data. (See, e.g., Meyers, supra, pp. 965-968.) Correlation between the location of dithp on a physical chromosomal map and a specific disorder, or a predisposition to a specific disorder, may help define the region of DNA associated with that disorder. The dithp sequences may also be used to detect polymorphisms that are genetically linked to the inheritance of a particular condition, disease, or disorder.

In situ hybridization of chromosomal preparations and genetic mapping techniques, such as linkage analysis using established chromosomal markers, may be used for extending existing genetic maps. Often the placement of a gene on the chromosome of another mammalian species, such as mouse, may reveal associated markers even if the number or arm of the corresponding human chromosome is not known. These new marker sequences can be mapped to human chromosomes and may provide valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once a disease or syndrome has been crudely correlated by genetic linkage with a particular genomic region, e.g., ataxia-telangiectasia to 11q22-23, any sequences mapping to that area may represent associated or regulatory genes for further investigation. (See, e.g., Gatti, R.A. et al. (1988) Nature 336:577-580.) The nucleotide sequences of the subject invention may also be used to detect differences in chromosomal architecture due to translocation, inversion, etc., among normal, carrier, or affected individuals.

Once a disease-associated gene is mapped to a chromosomal region, the gene must be cloned in order to identify mutations or other alterations (e.g., translocations or inversions) that may be correlated with disease. This process requires a physical map of the chromosomal region containing the disease-gene of interest along with associated markers. A physical map is necessary for determining the nucleotide sequence of and order of marker genes on a particular chromosomal region. Physical mapping techniques are well known in the art and require the generation of overlapping sets of cloned DNA fragments from a particular organelle, chromosome, or genome. These clones are analyzed to reconstruct and catalog their order. Once the position of a marker is determined, the DNA from that region is obtained by consulting the catalog and selecting clones from that region. The gene of interest is located through positional cloning techniques using hybridization or similar methods.

Diagnostic Uses

The dithp of the present invention may be used to design probes useful in diagnostic assays. Such assays, well known to those skilled in the art, may be used to detect or confirm conditions, disorders, or diseases associated with abnormal levels of dithp expression. Labeled probes developed from dithp sequences are added to a sample under hybridizing conditions of desired stringency. In some instances, dithp, or fragments or oligonucleotides derived from dithp, may be used as primers in amplification steps prior to hybridization. The amount of hybridization complex formed is quantified and compared with standards for that cell or tissue. If dithp expression varies significantly from the standard, the assay indicates the presence of the condition, disorder, or disease. Qualitative or quantitative diagnostic methods may include northern, dot blot, or other membrane or dip-stick based technologies or multiple-sample format technologies such as PCR, enzyme-linked immunosorbent assay (ELISA)-like, pin, or chip-based assays.

The probes described above may also be used to monitor the progress of conditions, disorders, or diseases associated with abnormal levels of dithp expression, or to evaluate the efficacy of a particular therapeutic treatment. The candidate probe may be identified from the dithp that are specific to a given human tissue and have not been observed in GenBank or other genome databases. Such a probe may be used in animal studies, preclinical tests, clinical trials, or in monitoring the treatment of an individual patient. In a typical process, standard expression is established by methods well known in the art for use as a basis of comparison, samples from patients affected by the disorder or disease are combined with the probe to evaluate any deviation from the standard profile, and a therapeutic agent is administered and effects are monitored to generate a treatment profile. Efficacy is evaluated by determining whether the expression progresses toward or returns to the standard normal pattern. Treatment profiles may be generated over a period of several days or several months. Statistical methods well known to those skilled in the art may be use to determine the significance of such

therapeutic agents.

The polynucleotides are also useful for identifying individuals from minute biological samples, for example, by matching the RFLP pattern of a sample's DNA to that of an individual's DNA. The polynucleotides of the present invention can also be used to determine the actual base-by-base DNA sequence of selected portions of an individual's genome. These sequences can be used to prepare PCR primers for amplifying and isolating such selected DNA, which can then be sequenced. Using this technique, an individual can be identified through a unique set of DNA sequences. Once a unique ID database is established for an individual, positive identification of that individual can be made from extremely small tissue samples.

In a particular aspect, oligonucleotide primers derived from the dithp of the invention may be used to detect single nucleotide polymorphisms (SNPs). SNPs are substitutions, insertions and deletions that are a frequent cause of inherited or acquired genetic disease in humans. Methods of SNP detection include, but are not limited to, single-stranded conformation polymorphism (SSCP) and fluorescent SSCP (fSSCP) methods. In SSCP, oligonucleotide primers derived from the polynucleotide sequences encoding DITHP are used to amplify DNA using the polymerase chain reaction (PCR). The DNA may be derived, for example, from diseased or normal tissue, biopsy samples, bodily fluids, and the like. SNPs in the DNA cause differences in the secondary and tertiary structures of PCR products in single-stranded form, and these differences are detectable using gel electrophoresis in non-denaturing gels. In fSCCP, the oligonucleotide primers are fluorescently labeled, which allows detection of the amplimers in high-throughput equipment such as DNA sequencing machines. Additionally, sequence database analysis methods, termed in silico SNP (isSNP), are capable of identifying polymorphisms by comparing the sequences of individual overlapping DNA fragments which assemble into a common consensus sequence. These computer-based methods filter out sequence variations due to laboratory preparation of DNA and sequencing errors using statistical models and automated analyses of DNA sequence chromatograms. In the alternative, SNPs may be detected and characterized by mass spectrometry using, for example, the high throughput MASSARRAY system (Sequenom, Inc., San Diego CA).

DNA-based identification techniques are critical in forensic technology. DNA sequences taken from very small biological samples such as tissues, e.g., hair or skin, or body fluids, e.g., blood, saliva, semen, etc., can be amplified using, e.g., PCR, to identify individuals. (See, e.g., Erlich, H. (1992) PCR Technology, Freeman and Co., New York, NY). Similarly, polynucleotides of the present invention can be used as polymorphic markers.

There is also a need for reagents capable of identifying the source of a particular tissue. Appropriate reagents can comprise, for example, DNA probes or primers prepared from the sequences of the present invention that are specific for particular tissues. Panels of such reagents can identify

tissue by species and/or by organ type. In a similar fashion, these reagents can be used to screen tissue cultures for contamination.

The polynucleotides of the present invention can also be used as molecular weight markers on nucleic acid gels or Southern blots, as diagnostic probes for the presence of a specific mRNA in a particular cell type, in the creation of subtracted cDNA libraries which aid in the discovery of novel polynucleotides, in selection and synthesis of oligomers for attachment to an array or other support, and as an antigen to elicit an immune response.

Disease Model Systems Using dithp

The dithp of the invention or their mammalian homologs may be "knocked out" in an animal model system using homologous recombination in embryonic stem (ES) cells. Such techniques are well known in the art and are useful for the generation of animal models of human disease. (See, e.g., U.S. Patent Number 5,175,383 and U.S. Patent Number 5,767,337.) For example, mouse ES cells, such as the mouse 129/SvJ cell line, are derived from the early mouse embryo and grown in culture. The ES cells are transformed with a vector containing the gene of interest disrupted by a marker gene, e.g., the neomycin phosphotransferase gene (neo; Capecchi, M.R. (1989) Science 244:1288-1292). The vector integrates into the corresponding region of the host genome by homologous recombination. Alternatively, homologous recombination takes place using the Cre-loxP system to knockout a gene of interest in a tissue- or developmental stage-specific manner (Marth, J.D. (1996) Clin. Invest. 97:1999-2002; Wagner, K.U. et al. (1997) Nucleic Acids Res. 25:4323-4330). Transformed ES cells are identified and microinjected into mouse cell blastocysts such as those from the C57BL/6 mouse strain. The blastocysts are surgically transferred to pseudopregnant dams, and the resulting chimeric progeny are genotyped and bred to produce heterozygous or homozygous strains. Transgenic animals thus generated may be tested with potential therapeutic or toxic agents.

The dithp of the invention may also be manipulated in vitro in ES cells derived from human blastocysts. Human ES cells have the potential to differentiate into at least eight separate cell lineages including endoderm, mesoderm, and ectodermal cell types. These cell lineages differentiate into, for example, neural cells, hematopoietic lineages, and cardiomyocytes (Thomson, J.A. et al. (1998) Science 282:1145-1147).

The dithp of the invention can also be used to create "knockin" humanized animals (pigs) or transgenic animals (mice or rats) to model human disease. With knockin technology, a region of dithp is injected into animal ES cells, and the injected sequence integrates into the animal cell genome. Transformed cells are injected into blastulae, and the blastulae are implanted as described above. Transgenic progeny or inbred lines are studied and treated with potential pharmaceutical agents to obtain information on treatment of a human disease. Alternatively, a mammal inbred to overexpress

dithp, resulting, e.g., in the secretion of DITHP in its milk, may also serve as a convenient source of that protein (Janne, J. et al. (1998) Biotechnol. Annu. Rev. 4:55-74).

Screening Assays

DITHP encoded by polynucleotides of the present invention may be used to screen for molecules that bind to or are bound by the encoded polypeptides. The binding of the polypeptide and the molecule may activate (agonist), increase, inhibit (antagonist), or decrease activity of the polypeptide or the bound molecule. Examples of such molecules include antibodies, oligonucleotides, proteins (e.g., receptors), or small molecules.

Preferably, the molecule is closely related to the natural ligand of the polypeptide, e.g., a ligand or fragment thereof, a natural substrate, or a structural or functional mimetic. (See, Coligan et al., (1991) Current Protocols in Immunology 1(2): Chapter 5.) Similarly, the molecule can be closely related to the natural receptor to which the polypeptide binds, or to at least a fragment of the receptor, e.g., the active site. In either case, the molecule can be rationally designed using known techniques. Preferably, the screening for these molecules involves producing appropriate cells which express the polypeptide, either as a secreted protein or on the cell membrane. Preferred cells include cells from mammals, yeast, Drosophila, or E. coli. Cells expressing the polypeptide or cell membrane fractions which contain the expressed polypeptide are then contacted with a test compound and binding, stimulation, or inhibition of activity of either the polypeptide or the molecule is analyzed.

An assay may simply test binding of a candidate compound to the polypeptide, wherein binding is detected by a fluorophore, radioisotope, enzyme conjugate, or other detectable label. Alternatively, the assay may assess binding in the presence of a labeled competitor.

Additionally, the assay can be carried out using cell-free preparations, polypeptide/molecule affixed to a solid support, chemical libraries, or natural product mixtures. The assay may also simply comprise the steps of mixing a candidate compound with a solution containing a polypeptide, measuring polypeptide/molecule activity or binding, and comparing the polypeptide/molecule activity or binding to a standard.

Preferably, an ELISA assay using, e.g., a monoclonal or polyclonal antibody, can measure polypeptide level in a sample. The antibody can measure polypeptide level by either binding, directly or indirectly, to the polypeptide or by competing with the polypeptide for a substrate.

All of the above assays can be used in a diagnostic or prognostic context. The molecules discovered using these assays can be used to treat disease or to bring about a particular result in a patient (e.g., blood vessel growth) by activating or inhibiting the polypeptide/molecule. Moreover, the assays can discover agents which may inhibit or enhance the production of the polypeptide from suitably manipulated cells or tissues.

Transcript Imaging

Another embodiment relates to the use of dithp to develop a transcript image of a tissue or cell type. A transcript image is the collective pattern of gene expression by a particular tissue or cell type under given conditions and at a given time. This pattern of gene expression is defined by the number of expressed genes, their abundance, and their function. Thus the dithp of the present invention may be used to develop a transcript image of a tissue or cell type by hybridizing, preferably in a microarray format, the dithp of the present invention to the totality of transcripts or reverse transcripts of a tissue or cell type. The resultant transcript image would provide a profile of gene activity pertaining to human molecules for diagnostics and therapeutics.

Transcript images which profile dithp expression may be generated using transcripts isolated from tissues, cell lines, biopsies, or other biological samples. The transcript image may thus reflect dithp expression in vivo, as in the case of a tissue or biopsy sample, or in vitro, as in the case of a cell line. Transcript images may be used to profile dithp expression in distinct tissue types. This process can be used to determine the activity of human diagnostic and therapeutic molecules in a particular tissue type relative to this activity in a different tissue type. Transcript images may be used to generate a profile of dithp expression characteristic of diseased tissue. Transcript images of tissues before and after treatment may be used for diagnostic purposes, to monitor the progression of disease, and to monitor the efficacy of drug treatments for diseases which affect the activity of human diagnostic and therapeutic molecules.

Transcript images which profile dithp expression may also be used in conjunction with in vitro model systems and preclinical evaluation of pharmaceuticals. Transcript images of cell lines can be used to assess the activity of human diagnostic and therapeutic molecules and/or to identify cell lines that lack or misregulate this activity. Such cell lines may then be treated with pharmaceutical agents, and a transcript image following treatment may indicate the efficacy of these agents in restoring desired levels of this activity. A similar approach may be used to assess the toxicity of pharmaceutical agents as reflected by undesirable changes in the activity of human diagnostic and therapeutic molecules. Candidate pharmaceutical agents may be evaluated by comparing their associated transcript images with those of pharmaceutical agents of known effectiveness.

Antisense Molecules

The polynucleotides of the present invention are useful in antisense technology. Antisense technology or therapy relies on the modulation of expression of a target protein through the specific binding of an antisense sequence to a target sequence encoding the target protein or directing its expression. (See, e.g., Agrawal, S., ed. (1996) Antisense Therapeutics, Humana Press Inc., Totawa NJ; Alama, A. et al. (1997) Pharmacol. Res. 36(3):171-178; Crooke, S.T. (1997) Adv. Pharmacol.

40:1-49; Sharma, H.W. and R. Narayanan (1995) Bioessays 17(12):1055-1063; and Lavrosky, Y. et al. (1997) Biochem. Mol. Med. 62(1):11-22.) An antisense sequence is a polynucleotide sequence capable of specifically hybridizing to at least a portion of the target sequence. Antisense sequences bind to cellular mRNA and/or genomic DNA, affecting translation and/or transcription. Antisense sequences can be DNA, RNA, or nucleic acid mimics and analogs. (See, e.g., Rossi, J.J. et al. (1991) Antisense Res. Dev. 1(3):285-288; Lee, R. et al. (1998) Biochemistry 37(3):900-1010; Pardridge, W.M. et al. (1995) Proc. Natl. Acad. Sci. USA 92(12):5592-5596; and Nielsen, P. E. and Haaima, G. (1997) Chem. Soc. Rev. 96:73-78.) Typically, the binding which results in modulation of expression occurs through hybridization or binding of complementary base pairs. Antisense sequences can also bind to DNA duplexes through specific interactions in the major groove of the double helix.

The polynucleotides of the present invention and fragments thereof can be used as antisense sequences to modify the expression of the polypeptide encoded by dithp. The antisense sequences can be produced ex vivo, such as by using any of the ABI nucleic acid synthesizer series (PE Biosystems) or other automated systems known in the art. Antisense sequences can also be produced biologically, such as by transforming an appropriate host cell with an expression vector containing the sequence of interest. (See, e.g., Agrawal, supra.)

In therapeutic use, any gene delivery system suitable for introduction of the antisense sequences into appropriate target cells can be used. Antisense sequences can be delivered intracellularly in the form of an expression plasmid which, upon transcription, produces a sequence complementary to at least a portion of the cellular sequence encoding the target protein. (See, e.g., Slater, J.E., et al. (1998) J. Allergy Clin. Immunol. 102(3):469-475; and Scanlon, K.J., et al. (1995) 9(13):1288-1296.) Antisense sequences can also be introduced intracellularly through the use of viral vectors, such as retrovirus and adeno-associated virus vectors. (See, e.g., Miller, A.D. (1990) Blood 76:271; Ausubel, F.M. et al. (1995) Current Protocols in Molecular Biology, John Wiley & Sons, New York NY; Uckert, W. and W. Walther (1994) Pharmacol. Ther. 63(3):323-347.) Other gene delivery mechanisms include liposome-derived systems, artificial viral envelopes, and other systems known in the art. (See, e.g., Rossi, J.J. (1995) Br. Med. Bull. 51(1):217-225; Boado, R.J. et al. (1998) J. Pharm. Sci. 87(11):1308-1315; and Morris, M.C. et al. (1997) Nucleic Acids Res. 25(14):2730-2736.)

Expression

In order to express a biologically active DITHP, the nucleotide sequences encoding DITHP or fragments thereof may be inserted into an appropriate expression vector, i.e., a vector which contains the necessary elements for transcriptional and translational control of the inserted coding sequence in a suitable host. Methods which are well known to those skilled in the art may be used to construct expression vectors containing sequences encoding DITHP and appropriate transcriptional and

translational control elements. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. (See, e.g., Sambrook, supra, Chapters 4, 8, 16, and 17; and Ausubel, supra, Chapters 9, 10, 13, and 16.)

A variety of expression vector/host systems may be utilized to contain and express sequences encoding DITHP. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with viral expression vectors (e.g., baculovirus); plant cell systems transformed with viral expression vectors (e.g., cauliflower mosaic virus, CaMV, or tobacco mosaic virus, TMV) or with bacterial expression vectors (e.g., Ti or pBR322 plasmids); or animal (mammalian) cell systems. (See, e.g., Sambrook, supra; Ausubel, 1995, supra, Van Heeke, G. and S.M. Schuster (1989) J. Biol. Chem. 264:5503-5509; Bitter, G.A. et al. (1987) Methods Enzymol. 153:516-544; Scorer, C.A. et al. (1994) Bio/Technology 12:181-184; Engelhard, E.K. et al. (1994) Proc. Natl. Acad. Sci. USA 91:3224-3227; Sandig, V. et al. (1996) Hum. Gene Ther. 7:1937-1945; Takamatsu, N. (1987) EMBO J. 6:307-311; Coruzzi, G. et al. (1984) EMBO J. 3:1671-1680; Broglie, R. et al. (1984) Science 224:838-843; Winter, J. et al. (1991) Results Probl. Cell Differ. 17:85-105; The McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York NY, pp. 191-196; Logan, J. and T. Shenk (1984) Proc. Natl. Acad. Sci. USA 81:3655-3659; and Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355.) Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia viruses, or from various bacterial plasmids, may be used for delivery of nucleotide sequences to the targeted organ, tissue, or cell population. (See, e.g., Di Nicola, M. et al. (1998) Cancer Gen. Ther. 5(6):350-356; Yu, M. et al., (1993) Proc. Natl. Acad. Sci. USA 90(13):6340-6344; Buller, R.M. et al. (1985) Nature 317(6040):813-815; McGregor, D.P. et al. (1994) Mol. Immunol. 31(3):219-226; and Verma, I.M. and N. Somia (1997) Nature 389:239-242.) The invention is not limited by the host cell employed.

For long term production of recombinant proteins in mammalian systems, stable expression of DITHP in cell lines is preferred. For example, sequences encoding DITHP can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Any number of selection systems may be used to recover transformed cell lines. (See, e.g., Wigler, M. et al. (1977) Cell 11:223-232; Lowy, I. et al. (1980) Cell 22:817-823.; Wigler, M. et al. (1980) Proc. Natl. Acad. Sci. USA 77:3567-3570; Colbere-Garapin, F. et al. (1981) J. Mol. Biol. 150:1-14; Hartman, S.C. and R.C.Mulligan (1988) Proc. Natl. Acad. Sci. USA 85:8047-8051; Rhodes, C.A. (1995) Methods Mol. Biol. 55:121-131.)

Therapeutic Uses of dithp

The dithp of the invention may be used for somatic or germline gene therapy. Gene therapy may be performed to (i) correct a genetic deficiency (e.g., in the cases of severe combined immunodeficiency (SCID)-X1 disease characterized by X-linked inheritance (Cavazzana-Calvo, M. et al. (2000) Science 288:669-672), severe combined immunodeficiency syndrome associated with an inherited adenosine deaminase (ADA) deficiency (Blaese, R.M. et al. (1995) Science 270:475-480; Bordignon, C. et al. (1995) Science 270:470-475), cystic fibrosis (Zabner, J. et al. (1993) Cell 75:207-216; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:643-666; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:667-703), thalassemias, familial hypercholesterolemia, and hemophilia resulting from Factor VIII or Factor IX deficiencies (Crystal, R.G. (1995) Science 270:404-410; Verma, I.M. and Somia, N. (1997) Nature 389:239-242)), (ii) express a conditionally lethal gene product (e.g., in the case of cancers which result from unregulated cell proliferation), or (iii) express a protein which affords protection against intracellular parasites (e.g., against human retroviruses, such as human immunodeficiency virus (HIV) (Baltimore, D. (1988) Nature 335:395-396; Poeschla, E. et al. (1996) Proc. Natl. Acad. Sci. USA. 93:11395-11399), hepatitis B or C virus (HBV, HCV); fungal parasites, such as Candida albicans and Paracoccidioides brasiliensis; and protozoan parasites such as Plasmodium falciparum and Trypanosoma cruzi). In the case where a genetic deficiency in dithp expression or regulation causes disease, the expression of dithp from an appropriate population of transduced cells may alleviate the clinical manifestations caused by the genetic deficiency.

In a further embodiment of the invention, diseases or disorders caused by deficiencies in dithp are treated by constructing mammalian expression vectors comprising dithp and introducing these vectors by mechanical means into dithp-deficient cells. Mechanical transfer technologies for use with cells in vivo or ex vitro include (i) direct DNA microinjection into individual cells, (ii) ballistic gold particle delivery, (iii) liposome-mediated transfection, (iv) receptor-mediated gene transfer, and (v) the use of DNA transposons (Morgan, R.A. and Anderson, W.F. (1993) Annu. Rev. Biochem. 62:191-217; Ivics, Z. (1997) Cell 91:501-510; Boulay, J-L. and Récipon, H. (1998) Curr. Opin. Biotechnol. 9:445-450).

Expression vectors that may be effective for the expression of dithp include, but are not limited to, the PCDNA 3.1, EPITAG, PRCCMV2, PREP, PVAX vectors (Invitrogen, Carlsbad CA), PCMV-SCRIPT, PCMV-TAG, PEGSH/PERV (Stratagene, La Jolla CA), and PTET-OFF, PTET-ON, PTRE2, PTRE2-LUC, PTK-HYG (Clontech, Palo Alto CA). The dithp of the invention may be expressed using (i) a constitutively active promoter, (e.g., from cytomegalovirus (CMV), Rous sarcoma virus (RSV), SV40 virus, thymidine kinase (TK), or β-actin genes), (ii) an inducible promoter (e.g., the tetracycline-regulated promoter (Gossen, M. and Bujard, H. (1992) Proc. Natl. Acad. Sci. U.S.A. 89:5547-5551; Gossen, M. et al., (1995) Science 268:1766-1769; Rossi, F.M.V. and Blau, H.M. (1998) Curr. Opin. Biotechnol. 9:451-456), commercially available in the T-REX plasmid

(Invitrogen)); the ecdysone-inducible promoter (available in the plasmids PVGRXR and PIND; Invitrogen); the FK506/rapamycin inducible promoter; or the RU486/mifepristone inducible promoter (Rossi, F.M.V. and Blau, H.M. supra)), or (iii) a tissue-specific promoter or the native promoter of the endogenous gene encoding DITHP from a normal individual.

Commercially available liposome transformation kits (e.g., the PERFECT LIPID TRANSFECTION KIT, available from Invitrogen) allow one with ordinary skill in the art to deliver polynucleotides to target cells in culture and require minimal effort to optimize experimental parameters. In the alternative, transformation is performed using the calcium phosphate method (Graham, F.L. and Eb, A.J. (1973) Virology 52:456-467), or by electroporation (Neumann, E. et al. (1982) EMBO J. 1:841-845). The introduction of DNA to primary cells requires modification of these standardized mammalian transfection protocols.

In another embodiment of the invention, diseases or disorders caused by genetic defects with respect to dithp expression are treated by constructing a retrovirus vector consisting of (i) the dithp of the invention under the control of an independent promoter or the retrovirus long terminal repeat (LTR) promoter, (ii) appropriate RNA packaging signals, and (iii) a Rev-responsive element (RRE) along with additional retrovirus *cis*-acting RNA sequences and coding sequences required for efficient vector propagation. Retrovirus vectors (e.g., PFB and PFBNEO) are commercially available (Stratagene) and are based on published data (Riviere, I. et al. (1995) Proc. Natl. Acad. Sci. U.S.A. 92:6733-6737), incorporated by reference herein. The vector is propagated in an appropriate vector producing cell line (VPCL) that expresses an envelope gene with a tropism for receptors on the target cells or a promiscuous envelope protein such as VSVg (Armentano, D. et al. (1987) J. Virol. 61:1647-1650; Bender, M.A. et al. (1987) J. Virol. 61:1639-1646; Adam, M.A. and Miller, A.D. (1988) J. Virol. 62:3802-3806; Dull, T. et al. (1998) J. Virol. 72:8463-8471; Zufferey, R. et al. (1998) J. Virol. 72:9873-9880). U.S. Patent Number 5,910,434 to Rigg ("Method for obtaining retrovirus packaging cell lines producing high transducing efficiency retroviral supernatant") discloses a method for obtaining retrovirus packaging cell lines and is hereby incorporated by reference. Propagation of retrovirus vectors, transduction of a population of cells (e.g., $CD4^+$ T-cells), and the return of transduced cells to a patient are procedures well known to persons skilled in the art of gene therapy and have been well documented (Ranga, U. et al. (1997) J. Virol. 71:7020-7029; Bauer, G. et al. (1997) Blood 89:2259-2267; Bonyhadi, M.L. (1997) J. Virol. 71:4707-4716; Ranga, U. et al. (1998) Proc. Natl. Acad. Sci. U.S.A. 95:1201-1206; Su, L. (1997) Blood 89:2283-2290).

In the alternative, an adenovirus-based gene therapy delivery system is used to deliver dithp to cells which have one or more genetic abnormalities with respect to the expression of dithp. The construction and packaging of adenovirus-based vectors are well known to those with ordinary skill in the art. Replication defective adenovirus vectors have proven to be versatile for importing genes

encoding immunoregulatory proteins into intact islets in the pancreas (Csete, M.E. et al. (1995) Transplantation 27:263-268). Potentially useful adenoviral vectors are described in U.S. Patent Number 5,707,618 to Armentano ("Adenovirus vectors for gene therapy"), hereby incorporated by reference. For adenoviral vectors, see also Antinozzi, P.A. et al. (1999) Annu. Rev. Nutr. 19:511-544 and Verma, I.M. and Somia, N. (1997) Nature 18:389:239-242, both incorporated by reference herein.

In another alternative, a herpes-based, gene therapy delivery system is used to deliver dithp to target cells which have one or more genetic abnormalities with respect to the expression of dithp. The use of herpes simplex virus (HSV)-based vectors may be especially valuable for introducing dithp to cells of the central nervous system, for which HSV has a tropism. The construction and packaging of herpes-based vectors are well known to those with ordinary skill in the art. A replication-competent herpes simplex virus (HSV) type 1-based vector has been used to deliver a reporter gene to the eyes of primates (Liu, X. et al. (1999) Exp. Eye Res.169:385-395). The construction of a HSV-1 virus vector has also been disclosed in detail in U.S. Patent Number 5,804,413 to DeLuca ("Herpes simplex virus strains for gene transfer"), which is hereby incorporated by reference. U.S. Patent Number 5,804,413 teaches the use of recombinant HSV d92 which consists of a genome containing at least one exogenous gene to be transferred to a cell under the control of the appropriate promoter for purposes including human gene therapy. Also taught by this patent are the construction and use of recombinant HSV strains deleted for ICP4, ICP27 and ICP22. For HSV vectors, see also Goins, W. F. et al. 1999 J. Virol. 73:519-532 and Xu, H. et al., (1994) Dev. Biol. 163:152-161, hereby incorporated by reference. The manipulation of cloned herpesvirus sequences, the generation of recombinant virus following the transfection of multiple plasmids containing different segments of the large herpesvirus genomes, the growth and propagation of herpesvirus, and the infection of cells with herpesvirus are techniques well known to those of ordinary skill in the art.

In another alternative, an alphavirus (positive, single-stranded RNA virus) vector is used to deliver dithp to target cells. The biology of the prototypic alphavirus, Semliki Forest Virus (SFV), has been studied extensively and gene transfer vectors have been based on the SFV genome (Garoff, H. and Li, K-J. (1998) Curr. Opin. Biotech. 9:464-469). During alphavirus RNA replication, a subgenomic RNA is generated that normally encodes the viral capsid proteins. This subgenomic RNA replicates to higher levels than the full-length genomic RNA, resulting in the overproduction of capsid proteins relative to the viral proteins with enzymatic activity (e.g., protease and polymerase). Similarly, inserting dithp into the alphavirus genome in place of the capsid-coding region results in the production of a large number of dithp RNAs and the synthesis of high levels of DITHP in vector transduced cells. While alphavirus infection is typically associated with cell lysis within a few days, the ability to establish a persistent infection in hamster normal kidney cells (BHK-21) with a variant of Sindbis virus (SIN) indicates that the lytic replication of alphaviruses can be altered to suit the needs of the gene

therapy application (Dryga, S.A. et al. (1997) Virology 228:74-83). The wide host range of alphaviruses will allow the introduction of DITHP into a variety of cell types. The specific transduction of a subset of cells in a population may require the sorting of cells prior to transduction. The methods of manipulating infectious cDNA clones of alphaviruses, performing alphavirus cDNA and RNA transfections, and performing alphavirus infections, are well known to those with ordinary skill in the art.

Antibodies

Anti-DITHP antibodies may be used to analyze protein expression levels. Such antibodies include, but are not limited to, polyclonal, monoclonal, chimeric, single chain, and Fab fragments. For descriptions of and protocols of antibody technologies, see, e.g., Pound, J.D. (1998) Immunochemical Protocols, Humana Press, Totowa, NJ.

The amino acid sequence encoded by the dithp of the Sequence Listing may be analyzed by appropriate software (e.g., LASERGENE NAVIGATOR software, DNASTAR) to determine regions of high immunogenicity. The optimal sequences for immunization are selected from the C-terminus, the N-terminus, and those intervening, hydrophilic regions of the polypeptide which are likely to be exposed to the external environment when the polypeptide is in its natural conformation. Analysis used to select appropriate epitopes is also described by Ausubel (1997, supra, Chapter 11.7). Peptides used for antibody induction do not need to have biological activity; however, they must be antigenic. Peptides used to induce specific antibodies may have an amino acid sequence consisting of at five amino acids, preferably at least 10 amino acids, and most preferably 15 amino acids. A peptide which mimics an antigenic fragment of the natural polypeptide may be fused with another protein such as keyhole limpet cyanin (KLH; Sigma, St. Louis MO) for antibody production. A peptide encompassing an antigenic region may be expressed from a dithp, synthesized as described above, or purified from human cells.

Procedures well known in the art may be used for the production of antibodies. Various hosts including mice, goats, and rabbits, may be immunized by injection with a peptide. Depending on the host species, various adjuvants may be used to increase immunological response.

In one procedure, peptides about 15 residues in length may be synthesized using an ABI 431A peptide synthesizer (PE Biosystems) using fmoc-chemistry and coupled to KLH (Sigma) by reaction with M-maleimidobenzoyl-N-hydroxysuccinimide ester (Ausubel, 1995, supra). Rabbits are immunized with the peptide-KLH complex in complete Freund's adjuvant. The resulting antisera are tested for antipeptide activity by binding the peptide to plastic, blocking with 1% bovine serum albumin (BSA), reacting with rabbit antisera, washing, and reacting with radioiodinated goat anti-rabbit IgG. Antisera with antipeptide activity are tested for anti-DITHP activity using protocols well known in the art, including ELISA, radioimmunoassay (RIA), and immunoblotting.

In another procedure, isolated and purified peptide may be used to immunize mice (about 100 μg of peptide) or rabbits (about 1 mg of peptide). Subsequently, the peptide is radioiodinated and used to screen the immunized animals' B-lymphocytes for production of antipeptide antibodies. Positive cells are then used to produce hybridomas using standard techniques. About 20 mg of peptide is sufficient for labeling and screening several thousand clones. Hybridomas of interest are detected by screening with radioiodinated peptide to identify those fusions producing peptide-specific monoclonal antibody. In a typical protocol, wells of a multi-well plate (FAST, Becton-Dickinson, Palo Alto, CA) are coated with affinity-purified, specific rabbit-anti-mouse (or suitable anti-species IgG) antibodies at 10 mg/ml. The coated wells are blocked with 1% BSA and washed and exposed to supernatants from hybridomas. After incubation, the wells are exposed to radiolabeled peptide at 1 mg/ml.

Clones producing antibodies bind a quantity of labeled peptide that is detectable above background. Such clones are expanded and subjected to 2 cycles of cloning. Cloned hybridomas are injected into pristane-treated mice to produce ascites, and monoclonal antibody is purified from the ascitic fluid by affinity chromatography on protein A (Amersham Pharmacia Biotech). Several procedures for the production of monoclonal antibodies, including in vitro production, are described in Pound (supra). Monoclonal antibodies with antipeptide activity are tested for anti-DITHP activity using protocols well known in the art, including ELISA, RIA, and immunoblotting.

Antibody fragments containing specific binding sites for an epitope may also be generated. For example, such fragments include, but are not limited to, the F(ab')2 fragments produced by pepsin digestion of the antibody molecule, and the Fab fragments generated by reducing the disulfide bridges of the F(ab')2 fragments. Alternatively, construction of Fab expression libraries in filamentous bacteriophage allows rapid and easy identification of monoclonal fragments with desired specificity (Pound, supra, Chaps. 45-47). Antibodies generated against polypeptide encoded by dithp can be used to purify and characterize full-length DITHP protein and its activity, binding partners, etc.

Assays Using Antibodies

Anti-DITHP antibodies may be used in assays to quantify the amount of DITHP found in a particular human cell. Such assays include methods utilizing the antibody and a label to detect expression level under normal or disease conditions. The peptides and antibodies of the invention may be used with or without modification or labeled by joining them, either covalently or noncovalently, with a reporter molecule.

Protocols for detecting and measuring protein expression using either polyclonal or monoclonal antibodies are well known in the art. Examples include ELISA, RIA, and fluorescent activated cell sorting (FACS). Such immunoassays typically involve the formation of complexes between the DITHP and its specific antibody and the measurement of such complexes. These and other assays are described

in Pound (supra).

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following preferred specific embodiments are, therefore, to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever.

The disclosures of all patents, applications, and publications mentioned above and below, in particular U.S. Ser. No. 60/137,109, U.S. Ser. No. 60/137,337, U.S. Ser. No. 60/137,258, U.S. Ser. No. 60/137,260, U.S. Ser. No. 60/137,113, U.S. Ser. No. 60/137,161, U.S. Ser. No. 60/137,417, U.S. Ser. No. 60/137,259, U.S. Ser. No. 60/137,396, U.S. Ser. No. 60/137,114, U.S. Ser. No. 60/137,173, U.S. Ser. No. 60/137,411, U.S. Ser. No. 60/147,436, U.S. Ser. No. 60/147,549, U.S. Ser. No. 60/147,377, U.S. Ser. No. 60/147,527, U.S. Ser. No. 60/147,520, U.S. Ser. No. 60/147,536, U.S. Ser. No. 60/147,530, U.S. Ser. No. 60/147,547, U.S. Ser. No. 60/147,824, U.S. Ser. No. 60/147,541, U.S. Ser. No. 60/147,542, and U.S. Ser. No. 60/147,500, are hereby expressly incorporated by reference.

## EXAMPLES

### I. Construction of cDNA Libraries

RNA was purchased from CLONTECH Laboratories, Inc. (Palo Alto CA) or isolated from various tissues. Some tissues were homogenized and lysed in guanidinium isothiocyanate, while others were homogenized and lysed in phenol or in a suitable mixture of denaturants, such as TRIZOL (Life Technologies), a monophasic solution of phenol and guanidine isothiocyanate. The resulting lysates were centrifuged over CsCl cushions or extracted with chloroform. RNA was precipitated with either isopropanol or sodium acetate and ethanol, or by other routine methods.

Phenol extraction and precipitation of RNA were repeated as necessary to increase RNA purity. In most cases, RNA was treated with DNase. For most libraries, poly(A+) RNA was isolated using oligo d(T)-coupled paramagnetic particles (Promega Corporation (Promega), Madison WI), OLIGOTEX latex particles (QIAGEN, Inc. (QIAGEN), Valencia CA), or an OLIGOTEX mRNA purification kit (QIAGEN). Alternatively, RNA was isolated directly from tissue lysates using other RNA isolation kits, e.g., the POLY(A)PURE mRNA purification kit (Ambion, Inc., Austin TX).

In some cases, Stratagene was provided with RNA and constructed the corresponding cDNA libraries. Otherwise, cDNA was synthesized and cDNA libraries were constructed with the UNIZAP vector system (Stratagene Cloning Systems, Inc. (Stratagene), La Jolla CA) or SUPERSCRIPT plasmid system (Life Technologies), using the recommended procedures or similar methods known in the art. (See, e.g., Ausubel, 1997, supra, Chapters 5.1 through 6.6.) Reverse transcription was initiated using oligo d(T) or random primers. Synthetic oligonucleotide adapters were ligated to double stranded cDNA, and the cDNA was digested with the appropriate restriction enzyme or enzymes. For

most libraries, the cDNA was size-selected (300-1000 bp) using SEPHACRYL S1000, SEPHAROSE CL2B, or SEPHAROSE CL4B column chromatography (Amersham Pharmacia Biotech) or preparative agarose gel electrophoresis. cDNAs were ligated into compatible restriction enzyme sites of the polylinker of a suitable plasmid, e.g., PBLUESCRIPT plasmid (Stratagene), pSPORT1 plasmid (Life Technologies), or pINCY (Incyte). Recombinant plasmids were transformed into competent E. coli cells including XL1-Blue, XL1-BlueMRF, or SOLR from Stratagene or DH5α, DH10B, or ElectroMAX DH10B from Life Technologies.

## II. Isolation of cDNA Clones

Plasmids were recovered from host cells by in vivo excision using the UNIZAP vector system (Stratagene) or by cell lysis. Plasmids were purified using at least one of the following: the Magic or WIZARD Minipreps DNA purification system (Promega); the AGTC Miniprep purification kit (Edge BioSystems, Gaithersburg MD); and the QIAWELL 8, QIAWELL 8 Plus, and QIAWELL 8 Ultra plasmid purification systems or the R.E.A.L. PREP 96 plasmid purification kit (QIAGEN). Following precipitation, plasmids were resuspended in 0.1 ml of distilled water and stored, with or without lyophilization, at 4°C.

Alternatively, plasmid DNA was amplified from host cell lysates using direct link PCR in a high-throughput format. (Rao, V.B. (1994) Anal. Biochem. 216:1-14.) Host cell lysis and thermal cycling steps were carried out in a single reaction mixture. Samples were processed and stored in 384-well plates, and the concentration of amplified plasmid DNA was quantified fluorometrically using PICOGREEN dye (Molecular Probes, Inc. (Molecular Probes), Eugene OR) and a FLUOROSKAN II fluorescence scanner (Labsystems Oy, Helsinki, Finland).

## III. Sequencing and Analysis

cDNA sequencing reactions were processed using standard methods or high-throughput instrumentation such as the ABI CATALYST 800 thermal cycler (PE Biosystems) or the PTC-200 thermal cycler (MJ Research) in conjunction with the HYDRA microdispenser (Robbins Scientific Corp., Sunnyvale CA) or the MICROLAB 2200 liquid transfer system (Hamilton). cDNA sequencing reactions were prepared using reagents provided by Amersham Pharmacia Biotech or supplied in ABI sequencing kits such as the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (PE Biosystems). Electrophoretic separation of cDNA sequencing reactions and detection of labeled polynucleotides were carried out using the MEGABACE 1000 DNA sequencing system (Molecular Dynamics); the ABI PRISM 373 or 377 sequencing system (PE Biosystems) in conjunction with standard ABI protocols and base calling software; or other sequence analysis systems known in the art. Reading frames within the cDNA sequences were identified using standard methods (reviewed in

Ausubel, 1997, supra, Chapter 7.7). Some of the cDNA sequences were selected for extension using the techniques disclosed in Example VIII.

**IV. Assembly and Analysis of Sequences**

Component sequences from chromatograms were subject to PHRED analysis and assigned a quality score. The sequences having at least a required quality score were subject to various pre-processing editing pathways to eliminate, e.g., low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, bacterial contamination sequences, and sequences smaller than 50 base pairs. In particular, low-information sequences and repetitive elements (e.g., dinucleotide repeats, Alu repeats, etc.) were replaced by "n's", or masked, to prevent spurious matches.

Processed sequences were then subject to assembly procedures in which the sequences were assigned to gene bins (bins). Each sequence could only belong to one bin. Sequences in each gene bin were assembled to produce consensus sequences (templates). Subsequent new sequences were added to existing bins using BLASTn (v.1.4 WashU) and CROSSMATCH. Candidate pairs were identified as all BLAST hits having a quality score greater than or equal to 150. Alignments of at least 82% local identity were accepted into the bin. The component sequences from each bin were assembled using a version of PHRAP. Bins with several overlapping component sequences were assembled using DEEP PHRAP. The orientation (sense or antisense) of each assembled template was determined based on the number and orientation of its component sequences. Template sequences as disclosed in the sequence listing correspond to sense strand sequences (the "forward" reading frames), to the best determination. The complementary (antisense) strands are inherently disclosed herein. The component sequences which were used to assemble each template consensus sequence are listed in Table 4, along with their positions along the template nucleotide sequences.

Bins were compared against each other and those having local similarity of at least 82% were combined and reassembled. Reassembled bins having templates of insufficient overlap (less than 95% local identity) were re-split. Assembled templates were also subject to analysis by STITCHER/EXON MAPPER algorithms which analyze the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types or disease states, etc. These resulting bins were subject to several rounds of the above assembly procedures.

Once gene bins were generated based upon sequence alignments, bins were clone joined based upon clone information. If the 5' sequence of one clone was present in one bin and the 3' sequence from the same clone was present in a different bin, it was likely that the two bins actually belonged together in a single bin. The resulting combined bins underwent assembly procedures to regenerate the

consensus sequences.

The final assembled templates were subsequently annotated using the following procedure. Template sequences were analyzed using BLASTn (v2.0, NCBI) versus gbpri (GenBank version 116). "Hits" were defined as an exact match having from 95% local identity over 200 base pairs through 100% local identity over 100 base pairs, or a homolog match having an E-value, i.e. a probability score, of $\leq 1 \times 10^{-8}$. The hits were subject to frameshift FASTx versus GENPEPT (GenBank version 116). (See Table 5). In this analysis, a homolog match was defined as having an E-value of $\leq 1 \times 10^{-8}$. The assembly method used above was described in "System and Methods for Analyzing Biomolecular Sequences," U.S.S.N. 09/276,534, filed March 25, 1999, and the LIFESEQ Gold user manual (Incyte) both incorporated by reference herein.

Following assembly, template sequences were subjected to motif, BLAST, and functional analyses, and categorized in protein hierarchies using methods described in, e.g., "Database System Employing Protein Function Hierarchies for Viewing Biomolecular Sequence Data," U.S.S.N. 08/812,290, filed March 6, 1997; "Relational Database for Storing Biomolecule Information," U.S.S.N. 08/947,845, filed October 9, 1997; "Project-Based Full-Length Biomolecular Sequence Database," U.S.S.N. 08/811,758, filed March 6, 1997; and "Relational Database and System for Storing Information Relating to Biomolecular Sequences," U.S.S.N. 09/034,807, filed March 4, 1998, all of which are incorporated by reference herein.

The template sequences were further analyzed by translating each template in all three forward reading frames and searching each translation against the Pfam database of hidden Markov model-based protein families and domains using the HMMER software package (available to the public from Washington University School of Medicine, St. Louis MO). Regions of templates which, when translated, contain similarity to Pfam consensus sequences are reported in Table 2, along with descriptions of Pfam protein domains and families. Only those Pfam hits with an E-value of $\leq 1 \times 10^{-3}$ are reported. (See also World Wide Web site http://pfam.wustl.edu/ for detailed descriptions of Pfam protein domains and families.)

Additionally, the template sequences were translated in all three forward reading frames, and each translation was searched against hidden Markov models for signal peptide and transmembrane domains using the HMMER software package. Construction of hidden Markov models and their usage in sequence analysis has been described. (See, for example, Eddy, S.R. (1996) Curr. Opin. Str. Biol. 6:361-365.) Regions of templates which, when translated, contain similarity to signal peptide or transmembrane domain consensus sequences are reported in Table 3. Only those signal peptide or transmembrane hits with a cutoff score of 11 bits or greater are reported. A cutoff score of 11 bits or greater corresponds to at least about 91-94% true-positives in signal peptide prediction, and at least about 75% true-positives in transmembrane domain prediction.

The results of HMMER analysis as reported in Tables 2 and 3 may support the results of BLAST analysis as reported in Table 1 or may suggest alternative or additional properties of template-encoded polypeptides not previously uncovered by BLAST or other analyses.

Template sequences are further analyzed using the bioinformatics tools listed in Table 5, or using sequence analysis software known in the art such as MACDNASIS PRO software (Hitachi Software Engineering, South San Francisco CA) and LASERGENE software (DNASTAR). Template sequences may be further queried against public databases such as the GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases.

## V. Analysis of Polynucleotide Expression

Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound. (See, e.g., Sambrook, supra, ch. 7; Ausubel, 1995, supra, ch. 4 and 16.)

Analogous computer techniques applying BLAST were used to search for identical or related molecules in cDNA databases such as GenBank or LIFESEQ (Incyte Pharmaceuticals). This analysis is much faster than multiple membrane-based hybridizations. In addition, the sensitivity of the computer search can be modified to determine whether any particular match is categorized as exact or similar. The basis of the search is the product score, which is defined as:

$$\frac{\text{BLAST Score} \times \text{Percent Identity}}{5 \times \text{minimum}\{\text{length(Seq. 1)}, \text{length(Seq. 2)}\}}$$

The product score takes into account both the degree of similarity between two sequences and the length of the sequence match. The product score is a normalized value between 0 and 100, and is calculated as follows: the BLAST score is multiplied by the percent nucleotide identity and the product is divided by (5 times the length of the shorter of the two sequences). The BLAST score is calculated by assigning a score of +5 for every base that matches in a high-scoring segment pair (HSP), and -4 for every mismatch. Two sequences may share more than one HSP (separated by gaps). If there is more than one HSP, then the pair with the highest BLAST score is used to calculate the product score. The product score represents a balance between fractional overlap and quality in a BLAST alignment. For example, a product score of 100 is produced only for 100% identity over the entire length of the shorter of the two sequences being compared. A product score of 70 is produced either by 100% identity and 70% overlap at one end, or by 88% identity and 100% overlap at the other. A product score of 50 is produced either by 100% identity and 50% overlap at one end, or 79% identity and 100% overlap.

## VI. Tissue Distribution Profiling

A tissue distribution profile is determined for each template by compiling the cDNA library tissue classifications of its component cDNA sequences. Each component sequence, is derived from a cDNA library constructed from a human tissue. Each human tissue is classified into one of the following categories: cardiovascular system; connective tissue; digestive system; embryonic structures; endocrine system; exocrine glands; genitalia, female; genitalia, male; germ cells; hemic and immune system; liver; musculoskeletal system; nervous system; pancreas; respiratory system; sense organs; skin; stomatognathic system; unclassified/mixed; or urinary tract. Template sequences, component sequences, and cDNA library/tissue information are found in the LIFESEQ GOLD database (Incyte Genomics, Palo Alto CA).

## VII. Transcript Image Analysis

Transcript images are generated as described in Seilhamer et al., "Comparative Gene Transcript Analysis," U.S. Patent Number 5,840,484, incorporated herein by reference.

## VIII. Extension of Polynucleotide Sequences and Isolation of a Full-length cDNA

Oligonucleotide primers designed using a dithp of the Sequence Listing are used to extend the nucleic acid sequence. One primer is synthesized to initiate 5' extension of the template, and the other primer, to initiate 3' extension of the template. The initial primers may be designed using OLIGO 4.06 software (National Biosciences, Inc. (National Biosciences), Plymouth MN), or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68°C to about 72°C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations are avoided. Selected human cDNA libraries are used to extend the sequence. If more than one extension is necessary or desired, additional or nested sets of primers are designed.

High fidelity amplification is obtained by PCR using methods well known in the art. PCR is performed in 96-well plates using the PTC-200 thermal cycler (MJ Research). The reaction mix contains DNA template, 200 nmol of each primer, reaction buffer containing $Mg^{2+}$, $(NH_4)_2SO_4$, and ß-mercaptoethanol, Taq DNA polymerase (Amersham Pharmacia Biotech), ELONGASE enzyme (Life Technologies), and Pfu DNA polymerase (Stratagene), with the following parameters for primer pair PCI A and PCI B: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68°C, 5 min; Step 7: storage at 4°C. In the alternative, the parameters for primer pair T7 and SK+ are as follows: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 57°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times;

Step 6: 68°C, 5 min; Step 7: storage at 4°C.

The concentration of DNA in each well is determined by dispensing 100 µl PICOGREEN quantitation reagent (0.25% (v/v); Molecular Probes) dissolved in 1X Tris-EDTA (TE) and 0.5 µl of undiluted PCR product into each well of an opaque fluorimeter plate (Corning Incorporated (Corning), Corning NY), allowing the DNA to bind to the reagent. The plate is scanned in a FLUOROSKAN II (Labsystems Oy) to measure the fluorescence of the sample and to quantify the concentration of DNA. A 5 µl to 10 µl aliquot of the reaction mixture is analyzed by electrophoresis on a 1 % agarose mini-gel to determine which reactions are successful in extending the sequence.

The extended nucleotides are desalted and concentrated, transferred to 384-well plates, digested with CviJI cholera virus endonuclease (Molecular Biology Research, Madison WI), and sonicated or sheared prior to religation into pUC 18 vector (Amersham Pharmacia Biotech). For shotgun sequencing, the digested nucleotides are separated on low concentration (0.6 to 0.8%) agarose gels, fragments are excised, and agar digested with AGAR ACE (Promega). Extended clones are religated using T4 ligase (New England Biolabs, Inc., Beverly MA) into pUC 18 vector (Amersham Pharmacia Biotech), treated with Pfu DNA polymerase (Stratagene) to fill-in restriction site overhangs, and transfected into competent E. coli cells. Transformed cells are selected on antibiotic-containing media, individual colonies are picked and cultured overnight at 37°C in 384-well plates in LB/2x carbenicillin liquid media.

The cells are lysed, and DNA is amplified by PCR using Taq DNA polymerase (Amersham Pharmacia Biotech) and Pfu DNA polymerase (Stratagene) with the following parameters: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 72°C, 2 min; Step 5: steps 2, 3, and 4 repeated 29 times; Step 6: 72°C, 5 min; Step 7: storage at 4°C. DNA is quantified by PICOGREEN reagent (Molecular Probes) as described above. Samples with low DNA recoveries are reamplified using the same conditions as described above. Samples are diluted with 20% dimethysulfoxide (1:2, v/v), and sequenced using DYENAMIC energy transfer sequencing primers and the DYENAMIC DIRECT kit (Amersham Pharmacia Biotech) or the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (PE Biosystems).

In like manner, the dithp is used to obtain regulatory sequences (promoters, introns, and enhancers) using the procedure above, oligonucleotides designed for such extension, and an appropriate genomic library.

## IX. Labeling of Probes and Southern Hybridization Analyses

Hybridization probes derived from the dithp of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA. The labeling of probe nucleotides between 100 and 1000 nucleotides in length is specifically described, but essentially the same procedure may be used

with larger cDNA fragments. Probe sequences are labeled at room temperature for 30 minutes using a T4 polynucleotide kinase, $\gamma^{32}$P-ATP, and 0.5X One-Phor-All Plus (Amersham Pharmacia Biotech) buffer and purified using a ProbeQuant G-50 Microcolumn (Amersham Pharmacia Biotech). The probe mixture is diluted to $10^7$ dpm/µg/ml hybridization buffer and used in a typical membrane-based hybridization analysis.

The DNA is digested with a restriction endonuclease such as Eco RV and is electrophoresed through a 0.7% agarose gel. The DNA fragments are transferred from the agarose to nylon membrane (NYTRAN Plus, Schleicher & Schuell, Inc., Keene NH) using procedures specified by the manufacturer of the membrane. Prehybridization is carried out for three or more hours at 68°C, and hybridization is carried out overnight at 68°C. To remove non-specific signals, blots are sequentially washed at room temperature under increasingly stringent conditions, up to 0.1x saline sodium citrate (SSC) and 0.5% sodium dodecyl sulfate. After the blots are placed in a PHOSPHORIMAGER cassette (Molecular Dynamics) or are exposed to autoradiography film, hybridization patterns of standard and experimental lanes are compared. Essentially the same procedure is employed when screening RNA.

**X. Chromosome Mapping of dithp**

The cDNA sequences which were used to assemble SEQ ID NO:1-52 are compared with sequences from the Incyte LIFESEQ database and public domain databases using BLAST and other implementations of the Smith-Waterman algorithm. Sequences from these databases that match SEQ ID NO:1-52 are assembled into clusters of contiguous and overlapping sequences using assembly algorithms such as PHRAP (Table 5). Radiation hybrid and genetic mapping data available from public resources such as the Stanford Human Genome Center (SHGC), Whitehead Institute for Genome Research (WIGR), and Généthon are used to determine if any of the clustered sequences have been previously mapped. Inclusion of a mapped sequence in a cluster will result in the assignment of all sequences of that cluster, including its particular SEQ ID NO:, to that map location. The genetic map locations of SEQ ID NO:1-52 are described as ranges, or intervals, of human chromosomes. The map position of an interval, in centiMorgans, is measured relative to the terminus of the chromosome's p-arm. (The centiMorgan (cM) is a unit of measurement based on recombination frequencies between chromosomal markers. On average, 1 cM is roughly equivalent to 1 megabase (Mb) of DNA in humans, although this can vary widely due to hot and cold spots of recombination.) The cM distances are based on genetic markers mapped by Généthon which provide boundaries for radiation hybrid markers whose sequences were included in each of the clusters.

**XI. Microarray Analysis**

Probe Preparation from Tissue or Cell Samples

Total RNA is isolated from tissue samples using the guanidinium thiocyanate method and polyA$^+$ RNA is purified using the oligo (dT) cellulose method. Each polyA$^+$ RNA sample is reverse transcribed using MMLV reverse-transcriptase, 0.05 pg/µl oligo-dT primer (21mer), 1X first strand buffer, 0.03 units/µl RNase inhibitor, 500 µM dATP, 500 µM dGTP, 500 µM dTTP, 40 µM dCTP, 40 µM dCTP-Cy3 (BDS) or dCTP-Cy5 (Amersham Pharmacia Biotech). The reverse transcription reaction is performed in a 25 ml volume containing 200 ng polyA$^+$ RNA with GEMBRIGHT kits (Incyte). Specific control polyA$^+$ RNAs are synthesized by in vitro transcription from non-coding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, the control mRNAs at 0.002 ng, 0.02 ng, 0.2 ng, and 2 ng are diluted into reverse transcription reaction at ratios of 1:100,000, 1:10,000, 1:1000, 1:100 (w/w) to sample mRNA respectively. The control mRNAs are diluted into reverse transcription reaction at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, 25:1 (w/w) to sample mRNA differential expression patterns. After incubation at 37° C for 2 hr, each reaction sample (one with Cy3 and another with Cy5 labeling) is treated with 2.5 ml of 0.5M sodium hydroxide and incubated for 20 minutes at 85°C to the stop the reaction and degrade the RNA. Probes are purified using two successive CHROMA SPIN 30 gel filtration spin columns (CLONTECH Laboratories, Inc. (CLONTECH), Palo Alto CA) and after combining, both reaction samples are ethanol precipitated using 1 ml of glycogen (1 mg/ml), 60 ml sodium acetate, and 300 ml of 100% ethanol. The probe is then dried to completion using a SpeedVAC (Savant Instruments Inc., Holbrook NY) and resuspended in 14 µl 5X SSC/0.2% SDS.

Microarray Preparation

Sequences of the present invention are used to generate array elements. Each array element is amplified from bacterial cells containing vectors with cloned cDNA inserts. PCR amplification uses primers complementary to the vector sequences flanking the cDNA insert. Array elements are amplified in thirty cycles of PCR from an initial quantity of 1-2 ng to a final quantity greater than 5 µg. Amplified array elements are then purified using SEPHACRYL-400 (Amersham Pharmacia Biotech).

Purified array elements are immobilized on polymer-coated glass slides. Glass microscope slides (Corning) are cleaned by ultrasound in 0.1% SDS and acetone, with extensive distilled water washes between and after treatments. Glass slides are etched in 4% hydrofluoric acid (VWR Scientific Products Corporation (VWR), West Chester, PA), washed extensively in distilled water, and coated with 0.05% aminopropyl silane (Sigma) in 95% ethanol. Coated slides are cured in a 110°C oven.

Array elements are applied to the coated glass substrate using a procedure described in US Patent No. 5,807,522, incorporated herein by reference. 1 µl of the array element DNA, at an average concentration of 100 ng/µl, is loaded into the open capillary printing element by a high-speed robotic apparatus. The apparatus then deposits about 5 nl of array element sample per slide.

Microarrays are UV-crosslinked using a STRATALINKER UV-crosslinker (Stratagene). Microarrays are washed at room temperature once in 0.2% SDS and three times in distilled water. Non-specific binding sites are blocked by incubation of microarrays in 0.2% casein in phosphate buffered saline (PBS) (Tropix, Inc., Bedford, MA) for 30 minutes at 60°C followed by washes in 0.2% SDS and distilled water as before.

Hybridization

Hybridization reactions contain 9 µl of probe mixture consisting of 0.2 µg each of Cy3 and Cy5 labeled cDNA synthesis products in 5X SSC, 0.2% SDS hybridization buffer. The probe mixture is heated to 65°C for 5 minutes and is aliquoted onto the microarray surface and covered with an 1.8 $cm^2$ coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 µl of 5x SSC in a corner of the chamber. The chamber containing the arrays is incubated for about 6.5 hours at 60°C. The arrays are washed for 10 min at 45°C in a first wash buffer (1X SSC, 0.1% SDS), three times for 10 minutes each at 45°C in a second wash buffer (0.1X SSC), and dried.

Detection

Reporter-labeled hybridization complexes are detected with a microscope equipped with an Innova 70 mixed gas 10 W laser (Coherent, Inc., Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Inc., Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective. The 1.8 cm x 1.8 cm array used in the present example is scanned with a resolution of 20 micrometers.

In two separate scans, a mixed gas multiline laser excites the two fluorophores sequentially. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the array and the photomultiplier tubes are used to filter the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. Each array is typically scanned twice, one scan per fluorophore using the appropriate filters at the laser source, although the apparatus is capable of recording the spectra from both fluorophores simultaneously.

The sensitivity of the scans is typically calibrated using the signal intensity generated by a cDNA control species added to the probe mix at a known concentration. A specific location on the array contains a complementary DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000. When two probes from different

sources (e.g., representing test and control cells), each labeled with a different fluorophore, are hybridized to a single array for the purpose of identifying genes that are differentially expressed, the calibration is done by labeling samples of the calibrating cDNA with the two fluorophores and adding identical amounts of each to the hybridization mixture.

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Inc., Norwood, MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using each fluorophore's emission spectrum.

A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis is the GEMTOOLS gene expression analysis program (Incyte).

## XII. Complementary Nucleic Acids

Sequences complementary to the dithp are used to detect, decrease, or inhibit expression of the naturally occurring nucleotide. The use of oligonucleotides comprising from about 15 to 30 base pairs is typical in the art. However, smaller or larger sequence fragments can also be used. Appropriate oligonucleotides are designed from the dithp using OLIGO 4.06 software (National Biosciences) or other appropriate programs and are synthesized using methods standard in the art or ordered from a commercial supplier. To inhibit transcription, a complementary oligonucleotide is designed from the most unique 5' sequence and used to prevent transcription factor binding to the promoter sequence. To inhibit translation, a complementary oligonucleotide is designed to prevent ribosomal binding and processing of the transcript.

## XIII. Expression of DITHP

Expression and purification of DITHP is accomplished using bacterial or virus-based expression systems. For expression of DITHP in bacteria, cDNA is subcloned into an appropriate vector containing an antibiotic resistance gene and an inducible promoter that directs high levels of cDNA transcription. Examples of such promoters include, but are not limited to, the *trp-lac* (*tac*) hybrid promoter and the T5 or T7 bacteriophage promoter in conjunction with the *lac* operator regulatory element. Recombinant vectors are transformed into suitable bacterial hosts, e.g., BL21(DE3). Antibiotic resistant bacteria express DITHP upon induction with isopropyl beta-D-

thiogalactopyranoside (IPTG). Expression of DITHP in eukaryotic cells is achieved by infecting insect or mammalian cell lines with recombinant Autographica californica nuclear polyhedrosis virus (AcMNPV), commonly known as baculovirus. The nonessential polyhedrin gene of baculovirus is replaced with cDNA encoding DITHP by either homologous recombination or bacterial-mediated transposition involving transfer plasmid intermediates. Viral infectivity is maintained and the strong polyhedrin promoter drives high levels of cDNA transcription. Recombinant baculovirus is used to infect Spodoptera frugiperda (Sf9) insect cells in most cases, or human hepatocytes, in some cases. Infection of the latter requires additional genetic modifications to baculovirus. (See e.g., Engelhard, supra; and Sandig, supra.)

In most expression systems, DITHP is synthesized as a fusion protein with, e.g., glutathione S-transferase (GST) or a peptide epitope tag, such as FLAG or 6-His, permitting rapid, single-step, affinity-based purification of recombinant fusion protein from crude cell lysates. GST, a 26-kilodalton enzyme from Schistosoma japonicum, enables the purification of fusion proteins on immobilized glutathione under conditions that maintain protein activity and antigenicity (Amersham Pharmacia Biotech). Following purification, the GST moiety can be proteolytically cleaved from DITHP at specifically engineered sites. FLAG, an 8-amino acid peptide, enables immunoaffinity purification using commercially available monoclonal and polyclonal anti-FLAG antibodies (Eastman Kodak Company, Rochester NY). 6-His, a stretch of six consecutive histidine residues, enables purification on metal-chelate resins (QIAGEN). Methods for protein expression and purification are discussed in Ausubel (1995, supra, Chapters 10 and 16). Purified DITHP obtained by these methods can be used directly in the following activity assay.

## XIV. Demonstration of DITHP Activity

DITHP activity is demonstrated through a variety of specific assays, some of which are outlined below.

Oxidoreductase activity of DITHP is measured by the increase in extinction coefficient of NAD(P)H coenzyme at 340 nm for the measurement of oxidation activity, or the decrease in extinction coefficient of NAD(P)H coenzyme at 340 nm for the measurement of reduction activity (Dalziel, K. (1963) J. Biol. Chem. 238:2850-2858). One of three substrates may be used: Asn-βGal, biocytidine, or ubiquinone-10. The respective subunits of the enzyme reaction, for example, cytochtome $c_1$-b oxidoreductase and cytochrome c, are reconstituted. The reaction mixture contains a)1-2 mg/ml DITHP; and b) 15 mM substrate, 2.4 mM $NAD(P)^+$ in 0.1 M phosphate buffer, pH 7.1 (oxidation reaction), or 2.0 mM NAD(P)H, in 0.1 M $Na_2HPO_4$ buffer, pH 7.4 ( reduction reaction); in a total volume of 0.1 ml. Changes in absorbance at 340 nm ($A_{340}$) are measured at 23.5° C using a recording spectrophotometer (Shimadzu Scientific Instruments, Inc., Pleasanton CA). The amount of NAD(P)H

is stoichiometrically equivalent to the amount of substrate initially present, and the change in $A_{340}$ is a direct measure of the amount of NAD(P)H produced; $\Delta A_{340}$ = 6620[NADH]. Oxidoreductase activity of DITHP activity is proportional to the amount of NAD(P)H present in the assay.

Transferase activity of DITHP is measured through assays such as a methyl transferase assay in which the transfer of radiolabeled methyl groups between a donor substrate and an acceptor substrate is measured (Bokar, J.A. et al. (1994) J. Biol. Chem. 269:17697-17704). Reaction mixtures (50 µl final volume) contain 15 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10 mM dithiothreitol, 3% polyvinylalcohol, 1.5 µCi [*methyl*-$^3$H]AdoMet (0.375 µM AdoMet) (DuPont-NEN), 0.6 µg DITHP, and acceptor substrate (0.4 µg [$^{35}$S]RNA or 6-mercaptopurine (6-MP) to 1 mM final concentration). Reaction mixtures are incubated at 30°C for 30 minutes, then 65°C for 5 minutes. The products are separated by chromatography or electrophoresis and the level of methyl transferase activity is determined by quantification of *methyl*-$^3$H recovery.

DITHP hydrolase activity is measured by the hydrolysis of appropriate synthetic peptide substrates conjugated with various chromogenic molecules in which the degree of hydrolysis is quantified by spectrophotometric (or fluorometric) absorption of the released chromophore. (Beynon, R.J. and J.S. Bond (1994) Proteolytic Enzymes: A Practical Approach, Oxford University Press, New York NY, pp. 25-55) Peptide substrates are designed according to the category of protease activity as endopeptidase (serine, cysteine, aspartic proteases), animopeptidase (leucine aminopeptidase), or carboxypeptidase (Carboxypeptidase A and B, procollagen C-proteinase).

DITHP isomerase activity such as peptidyl prolyl *cis/trans* isomerase activity can be assayed by an enzyme assay described by Rahfeld, J.U., et al. (1994) (FEBS Lett. 352: 180-184). The assay is performed at 10°C in 35 mM HEPES buffer, pH 7.8, containing chymotrypsin (0.5 mg/ml) and DITHP at a variety of concentrations. Under these assay conditions, the substrate, Suc-Ala-Xaa-Pro-Phe-4-NA, is in equilibrium with respect to the prolyl bond, with 80-95% in *trans* and 5-20% in *cis* conformation. An aliquot (2 ul) of the substrate dissolved in dimethyl sulfoxide (10 mg/ml) is added to the reaction mixture described above. Only the *cis* isomer of the substrate is a substrate for cleavage by chymotrypsin. Thus, as the substrate is isomerized by DITHP, the product is cleaved by chymotrypsin to produce 4-nitroanilide, which is detected by it's absorbance at 390 nm. 4-Nitroanilide appears in a time-dependent and a DITHP concentration-dependent manner.

An assay for DITHP activity associated with growth and development measures cell proliferation as the amount of newly initiated DNA synthesis in Swiss mouse 3T3 cells. A plasmid containing polynucleotides encoding DITHP is transfected into quiescent 3T3 cultured cells using methods well known in the art. The transiently transfected cells are then incubated in the presence of [$^3$H]thymidine, a radioactive DNA precursor. Where applicable, varying amounts of DITHP ligand are added to the transfected cells. Incorporation of [$^3$H]thymidine into acid-precipitable DNA is measured

over an appropriate time interval, and the amount incorporated is directly proportional to the amount of newly synthesized DNA.

Growth factor activity of DITHP is measured by the stimulation of DNA synthesis in Swiss mouse 3T3 cells (McKay, I. and I. Leigh, eds. (1993) Growth Factors: A Practical Approach, Oxford University Press, New York NY). Initiation of DNA synthesis indicates the cells' entry into the mitotic cycle and their commitment to undergo later division. 3T3 cells are competent to respond to most growth factors, not only those that are mitogenic, but also those that are involved in embryonic induction. This competence is possible because the in vivo specificity demonstrated by some growth factors is not necessarily inherent but is determined by the responding tissue. In this assay, varying amounts of DITHP are added to quiescent 3T3 cultured cells in the presence of [$^3$H]thymidine, a radioactive DNA precursor. DITHP for this assay can be obtained by recombinant means or from biochemical preparations. Incorporation of [$^3$H]thymidine into acid-precipitable DNA is measured over an appropriate time interval, and the amount incorporated is directly proportional to the amount of newly synthesized DNA. A linear dose-response curve over at least a hundred-fold DITHP concentration range is indicative of growth factor activity. One unit of activity per milliliter is defined as the concentration of DITHP producing a 50% response level, where 100% represents maximal incorporation of [$^3$H]thymidine into acid-precipitable DNA.

Alternatively, an assay for cytokine activity of DITHP measures the proliferation of leukocytes. In this assay, the amount of tritiated thymidine incorporated into newly synthesized DNA is used to estimate proliferative activity. Varying amounts of DITHP are added to cultured leukocytes, such as granulocytes, monocytes, or lymphocytes, in the presence of [$^3$H]thymidine, a radioactive DNA precursor. DITHP for this assay can be obtained by recombinant means or from biochemical preparations. Incorporation of [$^3$H]thymidine into acid-precipitable DNA is measured over an appropriate time interval, and the amount incorporated is directly proportional to the amount of newly synthesized DNA. A linear dose-response curve over at least a hundred-fold DITHP concentration range is indicative of DITHP activity. One unit of activity per milliliter is conventionally defined as the concentration of DITHP producing a 50% response level, where 100% represents maximal incorporation of [$^3$H]thymidine into acid-precipitable DNA.

An alternative assay for DITHP cytokine activity utilizes a Boyden micro chamber (Neuroprobe, Cabin John MD) to measure leukocyte chemotaxis (Vicari, supra). In this assay, about $10^5$ migratory cells such as macrophages or monocytes are placed in cell culture media in the upper compartment of the chamber. Varying dilutions of DITHP are placed in the lower compartment. The two compartments are separated by a 5 or 8 micron pore polycarbonate filter (Nucleopore, Pleasanton CA). After incubation at 37°C for 80 to 120 minutes, the filters are fixed in methanol and stained with appropriate labeling agents. Cells which migrate to the other side of the filter are counted using

standard microscopy. The chemotactic index is calculated by dividing the number of migratory cells counted when DITHP is present in the lower compartment by the number of migratory cells counted when only media is present in the lower compartment. The chemotactic index is proportional to the activity of DITHP.

Alternatively, cell lines or tissues transformed with a vector containing dithp can be assayed for DITHP activity by immunoblotting. Cells are denatured in SDS in the presence of β-mercaptoethanol, nucleic acids removed by ethanol precipitation, and proteins purified by acetone precipitation. Pellets are resuspended in 20 mM tris buffer at pH 7.5 and incubated with Protein G-Sepharose pre-coated with an antibody specific for DITHP. After washing, the Sepharose beads are boiled in electrophoresis sample buffer, and the eluted proteins subjected to SDS-PAGE. The SDS-PAGE is transferred to a nitrocellulose membrane for immunoblotting, and the DITHP activity is assessed by visualizing and quantifying bands on the blot using the antibody specific for DITHP as the primary antibody and $^{125}$I-labeled IgG specific for the primary antibody as the secondary antibody.

DITHP kinase activity is measured by phosphorylation of a protein substrate using γ-labeled [$^{32}$P]-ATP and quantitation of the incorporated radioactivity using a radioisotope counter. DITHP is incubated with the protein substrate, [$^{32}$P]-ATP, and an appropriate kinase buffer. The [$^{32}$P] incorporated into the product is separated from free [$^{32}$P]-ATP by electrophoresis and the incorporated [$^{32}$P] is counted. The amount of [$^{32}$P] recovered is proportional to the kinase activity of DITHP in the assay. A determination of the specific amino acid residue phosphorylated is made by phosphoamino acid analysis of the hydrolyzed protein.

In the alternative, DITHP activity is measured by the increase in cell proliferation resulting from transformation of a mammalian cell line such as COS7, HeLa or CHO with an eukaryotic expression vector encoding DITHP. Eukaryotic expression vectors are commercially available, and the techniques to introduce them into cells are well known to those skilled in the art. The cells are incubated for 48-72 hours after transformation under conditions appropriate for the cell line to allow expression of DITHP. Phase microscopy is then used to compare the mitotic index of transformed versus control cells. An increase in the mitotic index indicates DITHP activity.

In a further alternative, an assay for DITHP signaling activity is based upon the ability of GPCR family proteins to modulate G protein-activated second messenger signal transduction pathways (e.g., cAMP; Gaudin, P. et al. (1998) J. Biol. Chem. 273:4990-4996). A plasmid encoding full length DITHP is transfected into a mammalian cell line (e.g., Chinese hamster ovary (CHO) or human embryonic kidney (HEK-293) cell lines) using methods well-known in the art. Transfected cells are grown in 12-well trays in culture medium for 48 hours, then the culture medium is discarded, and the attached cells are gently washed with PBS. The cells are then incubated in culture medium with or without ligand for 30 minutes, then the medium is removed and cells lysed by treatment with

1 M perchloric acid. The cAMP levels in the lysate are measured by radioimmunoassay using methods well-known in the art. Changes in the levels of cAMP in the lysate from cells exposed to ligand compared to those without ligand are proportional to the amount of DITHP present in the transfected cells.

Alternatively, an assay for DITHP protein phosphatase activity measures the hydrolysis of P-nitrophenyl phosphate (PNPP). DITHP is incubated together with PNPP in HEPES buffer pH 7.5, in the presence of 0.1% β-mercaptoethanol at 37°C for 60 min. The reaction is stopped by the addition of 6 ml of 10 N NaOH, and the increase in light absorbance of the reaction mixture at 410 nm resulting from the hydrolysis of PNPP is measured using a spectrophotometer. The increase in light absorbance is proportional to the phosphatase activity of DITHP in the assay (Diamond, R.H. et al (1994) Mol Cell Biol 14:3752-3762).

An alternative assay measures DITHP-mediated G-protein signaling activity by monitoring the mobilization of $Ca^{++}$ as an indicator of the signal transduction pathway stimulation. (See, e.g., Grynkievicz, G. et al. (1985) J. Biol. Chem. 260:3440; McColl, S. et al. (1993) J. Immunol. 150:4550-4555; and Aussel, C. et al. (1988) J. Immunol. 140:215-220). The assay requires preloading neutrophils or T cells with a fluorescent dye such as FURA-2 or BCECF (Universal Imaging Corp, Westchester PA) whose emission characteristics are altered by $Ca^{++}$ binding. When the cells are exposed to one or more activating stimuli artificially (e.g., anti-CD3 antibody ligation of the T cell receptor) or physiologically (e.g., by allogeneic stimulation), $Ca^{++}$ flux takes place. This flux can be observed and quantified by assaying the cells in a fluorometer or fluorescent activated cell sorter. Measurements of $Ca^{++}$ flux are compared between cells in their normal state and those transfected with DITHP. Increased $Ca^{++}$ mobilization attributable to increased DITHP concentration is proportional to DITHP activity.

DITHP transport activity is assayed by measuring uptake of labeled substrates into Xenopus laevis oocytes. Oocytes at stages V and VI are injected with DITHP mRNA (10 ng per oocyte) and incubated for 3 days at 18°C in OR2 medium (82.5mM NaCl, 2.5 mM KCl, 1mM $CaCl_2$, 1mM $MgCl_2$, 1mM $Na_2HPO_4$, 5 mM Hepes, 3.8 mM NaOH, 50μg/ml gentamycin, pH 7.8) to allow expression of DITHP protein. Oocytes are then transferred to standard uptake medium (100mM NaCl, 2 mM KCl, 1mM $CaCl_2$, 1mM $MgCl_2$, 10 mM Hepes/Tris pH 7.5). Uptake of various substrates (e.g., amino acids, sugars, drugs, ions, and neurotransmitters) is initiated by adding labeled substrate (e.g. radiolabeled with $^3H$, fluorescently labeled with rhodamine, etc.) to the oocytes. After incubating for 30 minutes, uptake is terminated by washing the oocytes three times in $Na^+$-free medium, measuring the incorporated label, and comparing with controls. DITHP transport activity is proportional to the level of internalized labeled substrate.

DITHP transferase activity is demonstrated by a test for galactosyltransferase activity. This

can be determined by measuring the transfer of radiolabeled galactose from UDP-galactose to a GlcNAc-terminated oligosaccharide chain (Kolbinger, F. et al. (1998) J. Biol. Chem. 273:58-65). The sample is incubated with 14 µl of assay stock solution (180 mM sodium cacodylate, pH 6.5, 1 mg/ml bovine serum albumin, 0.26 mM UDP-galactose, 2 µl of UDP-[$^3$H]galactose), 1 µl of $MnCl_2$ (500 mM), and 2.5 µl of GlcNAcβO-$(CH_2)_8$-$CO_2Me$ (37 mg/ml in dimethyl sulfoxide) for 60 minutes at 37°C. The reaction is quenched by the addition of 1 ml of water and loaded on a C18 Sep-Pak cartridge (Waters), and the column is washed twice with 5 ml of water to remove unreacted UDP-[$^3$H]galactose. The [$^3$H]galactosylated GlcNAcβO-$(CH_2)_8$-$CO_2Me$ remains bound to the column during the water washes and is eluted with 5 ml of methanol. Radioactivity in the eluted material is measured by liquid scintillation counting and is proportional to galactosyltransferase activity in the starting sample.

In the alternative, DITHP induction by heat or toxins may be demonstrated using primary cultures of human fibroblasts or human cell lines such as CCL-13, HEK293, or HEP G2 (ATCC). To heat induce DITHP expression, aliquots of cells are incubated at 42 °C for 15, 30, or 60 minutes. Control aliquots are incubated at 37 °C for the same time periods. To induce DITHP expression by toxins, aliquots of cells are treated with 100 µM arsenite or 20 mM azetidine-2-carboxylic acid for 0, 3, 6, or 12 hours. After exposure to heat, arsenite, or the amino acid analogue, samples of the treated cells are harvested and cell lysates prepared for analysis by western blot. Cells are lysed in lysis buffer containing 1% Nonidet P-40, 0.15 M NaCl, 50 mM Tris-HCl, 5 mM EDTA, 2 mM N-ethylmaleimide, 2 mM phenylmethylsulfonyl fluoride, 1 mg/ml leupeptin, and 1 mg/ml pepstatin. Twenty micrograms of the cell lysate is separated on an 8% SDS-PAGE gel and transferred to a membrane. After blocking with 5% nonfat dry milk/phosphate-buffered saline for 1 h, the membrane is incubated overnight at 4°C or at room temperature for 2-4 hours with a 1:1000 dilution of anti-DITHP serum in 2% nonfat dry milk/phosphate-buffered saline. The membrane is then washed and incubated with a 1:1000 dilution of horseradish peroxidase-conjugated goat anti-rabbit IgG in 2% dry milk/phosphate-buffered saline. After washing with 0.1% Tween 20 in phosphate-buffered saline, the DITHP protein is detected and compared to controls using chemiluminescence.

Alternatively, DITHP protease activity is measured by the hydrolysis of appropriate synthetic peptide substrates conjugated with various chromogenic molecules in which the degree of hydrolysis is quantified by spectrophotometric (or fluorometric) absorption of the released chromophore (Beynon, R.J. and J.S. Bond (1994) Proteolytic Enzymes: A Practical Approach, Oxford University Press, New York, NY, pp.25-55). Peptide substrates are designed according to the category of protease activity as endopeptidase (serine, cysteine, aspartic proteases, or metalloproteases), aminopeptidase (leucine aminopeptidase), or carboxypeptidase (carboxypeptidases A and B, procollagen C-proteinase). Commonly used chromogens are 2-naphthylamine, 4-nitroaniline, and

furylacrylic acid. Assays are performed at ambient temperature and contain an aliquot of the enzyme and the appropriate substrate in a suitable buffer. Reactions are carried out in an optical cuvette, and the increase/decrease in absorbance of the chromogen released during hydrolysis of the peptide substrate is measured. The change in absorbance is proportional to the DITHP protease activity in the assay.

In the alternative, an assay for DITHP protease activity takes advantage of fluorescence resonance energy transfer (FRET) that occurs when one donor and one acceptor fluorophore with an appropriate spectral overlap are in close proximity. A flexible peptide linker containing a cleavage site specific for PRTS is fused between a red-shifted variant (RSGFP4) and a blue variant (BFP5) of Green Fluorescent Protein. This fusion protein has spectral properties that suggest energy transfer is occurring from BFP5 to RSGFP4. When the fusion protein is incubated with DITHP, the substrate is cleaved, and the two fluorescent proteins dissociate. This is accompanied by a marked decrease in energy transfer which is quantified by comparing the emission spectra before and after the addition of DITHP (Mitra, R.D. et al (1996) Gene 173:13-17). This assay can also be performed in living cells. In this case the fluorescent substrate protein is expressed constitutively in cells and DITHP is introduced on an inducible vector so that FRET can be monitored in the presence and absence of DITHP (Sagot, I. et al (1999) FEBS Lett. 447:53-57).

A method to determine the nucleic acid binding activity of DITHP involves a polyacrylamide gel mobility-shift assay. In preparation for this assay, DITHP is expressed by transforming a mammalian cell line such as COS7, HeLa or CHO with a eukaryotic expression vector containing DITHP cDNA. The cells are incubated for 48-72 hours after transformation under conditions appropriate for the cell line to allow expression and accumulation of DITHP. Extracts containing solubilized proteins can be prepared from cells expressing DITHP by methods well known in the art. Portions of the extract containing DITHP are added to [$^{32}$P]-labeled RNA or DNA. Radioactive nucleic acid can be synthesized in vitro by techniques well known in the art. The mixtures are incubated at 25°C in the presence of RNase- and DNase-inhibitors under buffered conditions for 5-10 minutes. After incubation, the samples are analyzed by polyacrylamide gel electrophoresis followed by autoradiography. The presence of a band on the autoradiogram indicates the formation of a complex between DITHP and the radioactive transcript. A band of similar mobility will not be present in samples prepared using control extracts prepared from untransformed cells.

In the alternative, a method to determine the methylase activity of a DITHP measures transfer of radiolabeled methyl groups between a donor substrate and an acceptor substrate. Reaction mixtures (50 µl final volume) contain 15 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10 mM dithiothreitol, 3% polyvinylalcohol, 1.5 µCi [*methyl*-$^3$H]AdoMet (0.375 µM AdoMet) (DuPont-NEN), 0.6 µg DITHP, and acceptor substrate (e.g., 0.4 µg [$^{35}$S]RNA, or 6-mercaptopurine (6-MP) to 1 mM final

concentration). Reaction mixtures are incubated at 30°C for 30 minutes, then 65°C for 5 minutes. Analysis of [*methyl*-$^3$H]RNA is as follows: 1) 50 µl of 2 x loading buffer (20 mM Tris-HCl, pH 7.6, 1 M LiCl, 1 mM EDTA, 1% sodium dodecyl sulphate (SDS)) and 50 µl oligo d(T)-cellulose (10 mg/ml in 1 x loading buffer) are added to the reaction mixture, and incubated at ambient temperature with shaking for 30 minutes. 2) Reaction mixtures are transferred to a 96-well filtration plate attached to a vacuum apparatus. 3) Each sample is washed sequentially with three 2.4 ml aliquots of 1 x oligo d(T) loading buffer containing 0.5% SDS, 0.1% SDS, or no SDS. and 4) RNA is eluted with 300 µl of water into a 96-well collection plate, transferred to scintillation vials containing liquid scintillant, and radioactivity determined. Analysis of [*methyl*-$^3$H]6-MP is as follows: 1) 500 µl 0.5 M borate buffer, pH 10.0, and then 2.5 ml of 20% (v/v) isoamyl alcohol in toluene are added to the reaction mixtures. 2) The samples mixed by vigorous vortexing for ten seconds. 3) After centrifugation at 700g for 10 minutes, 1.5 ml of the organic phase is transferred to scintillation vials containing 0.5 ml absolute ethanol and liquid scintillant, and radioactivity determined. and 4) Results are corrected for the extraction of 6-MP into the organic phase (approximately 41%).

An assay for adhesion activity of DITHP measures the disruption of cytoskeletal filament networks upon overexpression of DITHP in cultured cell lines (Rezniczek, G.A. et al. (1998) J. Cell Biol. 141:209-225). cDNA encoding DITHP is subcloned into a mammalian expression vector that drives high levels of cDNA expression. This construct is transfected into cultured cells, such as rat kangaroo PtK2 or rat bladder carcinoma 804G cells. Actin filaments and intermediate filaments such as keratin and vimentin are visualized by immunofluorescence microscopy using antibodies and techniques well known in the art. The configuration and abundance of cytoskeletal filaments can be assessed and quantified using confocal imaging techniques. In particular, the bundling and collapse of cytoskeletal filament networks is indicative of DITHP adhesion activity.

Alternatively, an assay for DITHP activity measures the expression of DITHP on the cell surface. cDNA encoding DITHP is transfected into a non-leukocytic cell line. Cell surface proteins are labeled with biotin (de la Fuente, M.A. et al. (1997) Blood 90:2398-2405). Immunoprecipitations are performed using DITHP-specific antibodies, and immunoprecipitated samples are analyzed using SDS-PAGE and immunoblotting techniques. The ratio of labeled immunoprecipitant to unlabeled immunoprecipitant is proportional to the amount of DITHP expressed on the cell surface.

Alternatively, an assay for DITHP activity measures the amount of cell aggregation induced by overexpression of DITHP. In this assay, cultured cells such as NIH3T3 are transfected with cDNA encoding DITHP contained within a suitable mammalian expression vector under control of a strong promoter. Cotransfection with cDNA encoding a fluorescent marker protein, such as Green Fluorescent Protein (CLONTECH), is useful for identifying stable transfectants. The amount of cell agglutination, or clumping, associated with transfected cells is compared with that associated with

untransfected cells. The amount of cell agglutination is a direct measure of DITHP activity.

DITHP may recognize and precipitate antigen from serum. This activity can be measured by the quantitative precipitin reaction (Golub, E.S. et al. (1987) Immunology: A Synthesis, Sinauer Associates, Sunderland MA, pages 113-115). DITHP is isotopically labeled using methods known in the art. Various serum concentrations are added to constant amounts of labeled DITHP. DITHP-antigen complexes precipitate out of solution and are collected by centrifugation. The amount of precipitable DITHP-antigen complex is proportional to the amount of radioisotope detected in the precipitate. The amount of precipitable DITHP-antigen complex is plotted against the serum concentration. For various serum concentrations, a characteristic precipitation curve is obtained, in which the amount of precipitable DITHP-antigen complex initially increases proportionately with increasing serum concentration, peaks at the equivalence point, and then decreases proportionately with further increases in serum concentration. Thus, the amount of precipitable DITHP-antigen complex is a measure of DITHP activity which is characterized by sensitivity to both limiting and excess quantities of antigen.

A microtubule motility assay for DITHP measures motor protein activity. In this assay, recombinant DITHP is immobilized onto a glass slide or similar substrate. Taxol-stabilized bovine brain microtubules (commercially available) in a solution containing ATP and cytosolic extract are perfused onto the slide. Movement of microtubules as driven by DITHP motor activity can be visualized and quantified using video-enhanced light microscopy and image analysis techniques. DITHP motor protein activity is directly proportional to the frequency and velocity of microtubule movement.

Alternatively, an assay for DITHP measures the formation of protein filaments in vitro. A solution of DITHP at a concentration greater than the "critical concentration" for polymer assembly is applied to carbon-coated grids. Appropriate nucleation sites may be supplied in the solution. The grids are negative stained with 0.7% (w/v) aqueous uranyl acetate and examined by electron microscopy. The appearance of filaments of approximately 25 nm (microtubules), 8 nm (actin), or 10 nm (intermediate filaments) is a demonstration of protein activity.

DITHP electron transfer activity is demonstrated by oxidation or reduction of NADP. Substrates such as Asn-βGal, biocytidine, or ubiquinone-10 may be used. The reaction mixture contains 1-2 mg/ml HORP, 15 mM substrate, and 2.4 mM $NAD(P)^+$ in 0.1 M phosphate buffer, pH 7.1 (oxidation reaction), or 2.0 mM NAD(P)H, in 0.1 M $Na_2HPO_4$ buffer, pH 7.4 (reduction reaction); in a total volume of 0.1 ml. FAD may be included with NAD, according to methods well known in the art. Changes in absorbance are measured using a recording spectrophotometer. The amount of NAD(P)H is stoichiometrically equivalent to the amount of substrate initially present, and the change in $A_{340}$ is a direct measure of the amount of NAD(P)H produced; $\Delta A_{340} = 6620[NADH]$. DITHP activity

is proportional to the amount of NAD(P)H present in the assay. The increase in extinction coefficient of NAD(P)H coenzyme at 340 nm is a measure of oxidation activity, or the decrease in extinction coefficient of NAD(P)H coenzyme at 340 nm is a measure of reduction activity (Dalziel, K. (1963) J. Biol. Chem. 238:2850-2858).

DITHP transcription factor activity is measured by its ability to stimulate transcription of a reporter gene (Liu, H.Y. et al. (1997) EMBO J. 16:5289-5298). The assay entails the use of a well characterized reporter gene construct, $LexA_{op}$-LacZ, that consists of LexA DNA transcriptional control elements ($LexA_{op}$) fused to sequences encoding the E. coli LacZ enzyme. The methods for constructing and expressing fusion genes, introducing them into cells, and measuring LacZ enzyme activity, are well known to those skilled in the art. Sequences encoding DITHP are cloned into a plasmid that directs the synthesis of a fusion protein, LexA-DITHP, consisting of DITHP and a DNA binding domain derived from the LexA transcription factor. The resulting plasmid, encoding a LexA-DITHP fusion protein, is introduced into yeast cells along with a plasmid containing the $LexA_{op}$-LacZ reporter gene. The amount of LacZ enzyme activity associated with LexA-DITHP transfected cells, relative to control cells, is proportional to the amount of transcription stimulated by the DITHP.

Chromatin activity of DITHP is demonstrated by measuring sensitivity to DNase I (Dawson, B.A. et al. (1989) J. Biol. Chem. 264:12830-12837). Samples are treated with DNase I, followed by insertion of a cleavable biotinylated nucleotide analog, 5-[(N-biotinamido)hexanoamido-ethyl-1,3-thiopropionyl-3-aminoallyl]-2'-deoxyuridine 5'-triphosphate using nick-repair techniques well known to those skilled in the art. Following purification and digestion with EcoRI restriction endonuclease, biotinylated sequences are affinity isolated by sequential binding to streptavidin and biotincellulose.

Another specific assay demonstrates the ion conductance capacity of DITHP using an electrophysiological assay. DITHP is expressed by transforming a mammalian cell line such as COS7, HeLa or CHO with a eukaryotic expression vector encoding DITHP. Eukaryotic expression vectors are commercially available, and the techniques to introduce them into cells are well known to those skilled in the art. A small amount of a second plasmid, which expresses any one of a number of marker genes such as β-galactosidase, is co-transformed into the cells in order to allow rapid identification of those cells which have taken up and expressed the foreign DNA. The cells are incubated for 48-72 hours after transformation under conditions appropriate for the cell line to allow expression and accumulation of DITHP and β-galactosidase. Transformed cells expressing β-galactosidase are stained blue when a suitable colorimetric substrate is added to the culture media under conditions that are well known in the art. Stained cells are tested for differences in membrane conductance due to various ions by electrophysiological techniques that are well known in the art. Untransformed cells, and/or cells transformed with either vector sequences alone or β-galactosidase sequences alone, are used as controls and tested in parallel. The contribution of DITHP to cation or

anion conductance can be shown by incubating the cells using antibodies specific for either DITHP. The respective antibodies will bind to the extracellular side of DITHP, thereby blocking the pore in the ion channel, and the associated conductance.

## XV. Functional Assays

DITHP function is assessed by expressing dithp at physiologically elevated levels in mammalian cell culture systems. cDNA is subcloned into a mammalian expression vector containing a strong promoter that drives high levels of cDNA expression. Vectors of choice include pCMV SPORT (Life Technologies) and pCR3.1 (Invitrogen Corporation, Carlsbad CA), both of which contain the cytomegalovirus promoter. 5-10 µg of recombinant vector are transiently transfected into a human cell line, preferably of endothelial or hematopoietic origin, using either liposome formulations or electroporation. 1-2 µg of an additional plasmid containing sequences encoding a marker protein are co-transfected.

Expression of a marker protein provides a means to distinguish transfected cells from nontransfected cells and is a reliable predictor of cDNA expression from the recombinant vector. Marker proteins of choice include, e.g., Green Fluorescent Protein (GFP; CLONTECH), CD64, or a CD64-GFP fusion protein. Flow cytometry (FCM), an automated laser optics-based technique, is used to identify transfected cells expressing GFP or CD64-GFP and to evaluate the apoptotic state of the cells and other cellular properties.

FCM detects and quantifies the uptake of fluorescent molecules that diagnose events preceding or coincident with cell death. These events include changes in nuclear DNA content as measured by staining of DNA with propidium iodide; changes in cell size and granularity as measured by forward light scatter and 90 degree side light scatter; down-regulation of DNA synthesis as measured by decrease in bromodeoxyuridine uptake; alterations in expression of cell surface and intracellular proteins as measured by reactivity with specific antibodies; and alterations in plasma membrane composition as measured by the binding of fluorescein-conjugated Annexin V protein to the cell surface. Methods in flow cytometry are discussed in Ormerod, M. G. (1994) Flow Cytometry, Oxford, New York NY.

The influence of DITHP on gene expression can be assessed using highly purified populations of cells transfected with sequences encoding DITHP and either CD64 or CD64-GFP. CD64 and CD64-GFP are expressed on the surface of transfected cells and bind to conserved regions of human immunoglobulin G (IgG). Transfected cells are efficiently separated from nontransfected cells using magnetic beads coated with either human IgG or antibody against CD64 (DYNAL, Inc., Lake Success NY). mRNA can be purified from the cells using methods well known by those of skill in the art. Expression of mRNA encoding DITHP and other genes of interest can be analyzed by northern analysis

or microarray techniques.

**XVI. Production of Antibodies**

DITHP substantially purified using polyacrylamide gel electrophoresis (PAGE; see, e.g., Harrington, M.G. (1990) Methods Enzymol. 182:488-495), or other purification techniques, is used to immunize rabbits and to produce antibodies using standard protocols.

Alternatively, the DITHP amino acid sequence is analyzed using LASERGENE software (DNASTAR) to determine regions of high immunogenicity, and a corresponding peptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. (See, e.g., Ausubel, 1995, supra, Chapter 11.)

Typically, peptides 15 residues in length are synthesized using an ABI 431A peptide synthesizer (PE Biosystems) using fmoc-chemistry and coupled to KLH (Sigma) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) to increase immunogenicity. (See, e.g., Ausubel, supra.) Rabbits are immunized with the peptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for antipeptide activity by, for example, binding the peptide to plastic, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG. Antisera with antipeptide activity are tested for anti-DITHP activity using protocols well known in the art, including ELISA, RIA, and immunoblotting.

**XVII. Purification of Naturally Occurring DITHP Using Specific Antibodies**

Naturally occurring or recombinant DITHP is substantially purified by immunoaffinity chromatography using antibodies specific for DITHP. An immunoaffinity column is constructed by covalently coupling anti-DITHP antibody to an activated chromatographic resin, such as CNBr-activated SEPHAROSE (Amersham Pharmacia Biotech). After the coupling, the resin is blocked and washed according to the manufacturer's instructions.

Media containing DITHP are passed over the immunoaffinity column, and the column is washed under conditions that allow the preferential absorbance of DITHP (e.g., high ionic strength buffers in the presence of detergent). The column is eluted under conditions that disrupt antibody/DITHP binding (e.g., a buffer of pH 2 to pH 3, or a high concentration of a chaotrope, such as urea or thiocyanate ion), and DITHP is collected.

**XVIII. Identification of Molecules Which Interact with DITHP**

DITHP, or biologically active fragments thereof, are labeled with $^{125}I$ Bolton-Hunter reagent. (See, e.g., Bolton, A.E. and W.M. Hunter (1973) Biochem. J. 133:529-539.) Candidate molecules

previously arrayed in the wells of a multi-well plate are incubated with the labeled DITHP, washed, and any wells with labeled DITHP complex are assayed. Data obtained using different concentrations of DITHP are used to calculate values for the number, affinity, and association of DITHP with the candidate molecules.

Alternatively, molecules interacting with DITHP are analyzed using the yeast two-hybrid system as described in Fields, S. and O. Song (1989) Nature 340:245-246, or using commercially available kits based on the two-hybrid system, such as the MATCHMAKER system (CLONTECH).

DITHP may also be used in the PATHCALLING process (CuraGen Corp., New Haven CT) which employs the yeast two-hybrid system in a high-throughput manner to determine all interactions between the proteins encoded by two large libraries of genes (Nandabalan, K. et al. (2000) U.S. Patent No. 6,057,101).

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

TABLE 1

| SEQ ID NO: | Template ID | GI Number | Probability Score | Annotation |
|---|---|---|---|---|
| 1 | 061149.1.j | g3073775 | 3.20E-31 | heparan-sulfate 6-sulfotransferase |
| 2 | 404508.3.j | g55535 | 1.30E-35 | 100 kDa protein |
| 3 | 441227.2.j | g2896147 | 0 | Human alpha-methylacyl-CoA racemase mRNA, complete cds. |
| 4 | 277927.2 | g2708639 | 2.80E-49 | carbonic anhydrase precursor |
| 5 | 475311.1 | g6815284 | 0 | Homo sapiens methionine adenosyltransferase regulatory beta subunit mRNA, complete cds. |
| 6 | 13039.2 | g35790 | 1.70E-12 | protein-tyrosine phosphatase |
| 7 | 238005.4 | g4321619 | 9.40E-96 | seven transmembrane domain orphan receptor |
| 8 | 345322.1.j | g206352 | 1.50E-15 | protein phosphatase inhibitor-1 protein |
| 9 | 348094.6.j | g3876117 | 3.80E-77 | similarity to mouse nek1 protein kinase |
| 10 | 233678.1 | g6179911 | 0 | Homo sapiens ADP-ribosylation factor binding protein GGA1 (GGA1) mRNA, complete cds. |
| 11 | 312243.1 | g3925265 | 6.90E-21 | similar to Probable rabGAP domains |
| 12 | 425487.3 | g6179913 | 1.00E-154 | Homo sapiens ADP-ribosylation factor binding protein GGA2 (GGA2) mRNA, complete cds. |
| 13 | 346813.3.j | g2088550 | 0 | Human hereditary haemochromatosis region, histone 2A-like protein gene, hereditary haemochromatosis (HLA-H) gene, RoRet gene, and sodium phosphate transporter (NPT3) gene, complete cds. |
| 14 | 006861.1.j | g1263080 | 4.00E-52 | Human mariner1 transposase gene, complete consensus sequence. |
| 15 | 028008.3.j | g2463648 | 5.00E-37 | snRNP core Sm protein homolog Sm-X5 |
| 16 | 346078.5.j | g3746839 | 0 | Human 45kDa splicing factor mRNA, complete cds. |
| 17 | 394637.1.j | g1263080 | 0 | Human mariner1 transposase gene, complete consensus sequence. |
| 18 | 222429.3 | g337451 | 1.90E-17 | hnRNP type A/B protein |
| 19 | 366739.2 | g5524926 | 0 | Homo sapiens mRNA for deoxyribonuclease III (drn3 gene). |
| 20 | 474635.6 | g3880433 | 4.90E-16 | similar to mitochondrial RNA splicing MSR4 like protein; cDNA EST EMBL:C09217 comes from this gene |
| 21 | 228470.1.j | g514373 | 1.00E-80 | Opioid-binding cell adhesion molecule |
| 22 | 407090.5.j | g500858 | 1.20E-10 | 50 kDa lectin |
| 23 | 068194.1.j | g397947 | 1.40E-23 | thioredoxin |
| 24 | 411449.2 | g5295993 | 0 | Homo sapiens SDHD gene for small subunit of cytochrome b of succinate dehydrogenase, complete cds. |

TABLE 1

| SEQ ID NO: | Template ID | GI Number | Probability Score | Annotation |
|---|---|---|---|---|
| 25 | 18549.2 | g4455608 | 1.50E-20 | dJ742C19.2 (APOBEC1 (Apolipoprotein B mRNA editing protein) and Phorbolin 1 LIKE) |
| 26 | 236043.3 | g562263 | 1.20E-94 | tropomodulin |
| 27 | 445433.2 | g28251 | 6.00E-25 | Human mRNA for beta-actin. |
| 28 | 344630.7.j | g40828 | 5.40E-30 | Hyd gamma (homologous to the periplasmic hydrogenase) |
| 29 | 257121.2 | g550123 | 2.30E-163 | GEG-154 |
| 30 | 243794.1.j | g36131 | 0 | Human mRNA for ribosomal protein L32. |
| 31 | 442085.1.j | g437878 | 4.50E-25 | ribosomal protein S24 |
| 32 | 370661.3.j | g1016712 | 1.40E-28 | Fos-related antigen |
| 33 | 427939.17.j | g488555 | 1.20E-85 | zinc finger protein ZNF135 |
| 34 | 430569.2.j | g487784 | 9.00E-16 | Human zinc finger protein ZNF136. |
| 35 | 444689.1.j | g487784 | 1.00E-26 | Human zinc finger protein ZNF136. |
| 36 | 445198.1.j | g498152 | 3.70E-35 | ha0946 protein is Kruppel-related |
| 37 | 84399.1 | g207696 | 2.80E-17 | zinc finger protein |
| 38 | 350044.1 | g1613847 | 5.00E-10 | Human zinc finger protein zfp6 (ZF6) mRNA, partial cds. |
| 39 | 441329.2 | g3006231 | 4.10E-78 | Zn-finger-like protein; similar to Z98745 (PID:g2924250) |
| 40 | 442401.2 | g498722 | 4.00E-62 | Human HZF2 mRNA for zinc finger protein. |
| 41 | 444933.2 | g456269 | 1.80E-20 | zinc finger protein 30 |
| 42 | 481129.4 | g3876716 | 3.10E-15 | similar to Zinc finger, C3HC4 type (RING finger) |
| 43 | 481999.1 | g4519269 | 3.00E-45 | Human ZK1 mRNA for Kruppel-type zinc finger protein, complete cds. |
| 44 | 233814.1.j | g1326367 | 7.50E-11 | Similar to the mitochondrial carrier family |
| 45 | 351376.4.j | g517226 | 7.40E-15 | mitochondrial ATPase inhibitor |
| 46 | 338992.1 | g414797 | 2.80E-253 | pyruvate dehydrogenase phosphatase |
| 47 | 200587.3.j | g722379 | 7.50E-23 | similar to NIFS protein (nitrogen fixation) |
| 48 | 246727.5.j | g337456 | 4.00E-25 | Human ribonucleoprotein (La) mRNA, 3' end. |
| 49 | 407087.3.j | g1176422 | 7.40E-72 | rhophilin |
| 50 | 441779.1.j | g1304381 | 8.10E-56 | hemoglobin alpha chain |
| 51 | 206603.1 | g4101720 | 4.00E-14 | lymphocyte specific formin related protein |
| 52 | 435694.2 | g4572570 | 3.00E-22 | Homo sapiens jun dimerization protein gene, partial cds; cfos gene, complete cds; and unknown gene. |

TABLE 2

| SEQ ID NO: | Template ID | Start | Stop | Frame | Pfam Hit | Pfam Description | E-value |
|---|---|---|---|---|---|---|---|
| 6 | 13039.2 | 568 | 801 | forward 1 | fn3 | Fibronectin type III domain | 3.00E-09 |
| 9 | 348094.6.j | 255 | 779 | forward 3 | pkinase | Eukaryotic protein kinase domain | 1.10E-53 |
| 9 | 348094.6.j | 782 | 967 | forward 2 | pkinase | Eukaryotic protein kinase domain | 2.70E-08 |
| 10 | 233678.1 | 149 | 514 | forward 2 | VHS | VHS domain | 1.30E-06 |
| 11 | 312243.1 | 703 | 1263 | forward 1 | TBC | TBC domain | 2.20E-07 |
| 12 | 425487.3 | 78 | 407 | forward 3 | VHS | VHS domain | 2.90E-09 |
| 16 | 346078.5.j | 844 | 978 | forward 1 | G-patch | G-patch domain | 8.00E-08 |
| 20 | 474635.6 | 258 | 530 | forward 3 | mito_carr | Mitochondrial carrier proteins | 4.10E-17 |
| 20 | 474635.6 | 2449 | 2652 | forward 1 | mito_carr | Mitochondrial carrier proteins | 2.40E-10 |
| 21 | 228470.1.j | 312 | 515 | forward 3 | ig | Immunoglobulin domain | 6.20E-08 |
| 29 | 257121.2 | 344 | 541 | forward 2 | IBR | IBR domain | 7.70E-23 |
| 30 | 243794.1.j | 456 | 740 | forward 3 | Ribosomal_L32e | Ribosomal protein L32 | 8.00E-20 |
| 30 | 243794.1.j | 231 | 326 | forward 3 | Ribosomal_S14 | Ribosomal protein S14p/S29e | 1.40E-12 |
| 31 | 442085.1.j | 149 | 400 | forward 2 | Ribosomal_S24e | Ribosomal protein S24e | 4.70E-40 |
| 33 | 427939.17.j | 1002 | 1070 | forward 3 | zf-C2H2 | Zinc finger, C2H2 type | 1.70E-07 |
| 33 | 427939.17.j | 589 | 657 | forward 1 | zf-C2H2 | Zinc finger, C2H2 type | 4.80E-07 |
| 35 | 444689.1.j | 170 | 274 | forward 2 | KRAB | KRAB box | 8.70E-06 |
| 36 | 445198.1.j | 80 | 268 | forward 2 | KRAB | KRAB box | 2.90E-42 |
| 38 | 350044.1 | 76 | 240 | forward 1 | KRAB | KRAB box | 4.40E-12 |
| 38 | 350044.1 | 895 | 963 | forward 1 | zf-C2H2 | Zinc finger, C2H2 type | 8.40E-06 |
| 39 | 441329.2 | 684 | 965 | forward 3 | SCAN | SCAN domain | 5.30E-48 |
| 40 | 442401.2 | 206 | 394 | forward 2 | KRAB | KRAB box | 2.40E-41 |
| 41 | 444933.2 | 292 | 480 | forward 1 | KRAB | KRAB box | 1.40E-41 |
| 43 | 481999.1 | 371 | 547 | forward 2 | KRAB | KRAB box | 1.90E-15 |
| 43 | 481999.1 | 839 | 907 | forward 2 | zf-C2H2 | Zinc finger, C2H2 type | 1.40E-04 |
| 44 | 233814.1.j | 356 | 625 | forward 2 | mito_carr | Mitochondrial carrier proteins | 2.10E-19 |
| 46 | 338992.1 | 1156 | 1680 | forward 1 | PP2C | Protein phosphatase 2C | 1.90E-54 |
| 46 | 338992.1 | 840 | 1184 | forward 3 | PP2C | Protein phosphatase 2C | 3.30E-10 |
| 47 | 200587.3.j | 146 | 1276 | forward 2 | aminotran_5 | Aminotransferases class-V | 3.20E-32 |
| 50 | 441779.1.j | 76 | 498 | forward 1 | globin | Globin | 4.20E-53 |

TABLE 3

| SEQ ID NO: | Template ID | Start | Stop | Frame | Domain Type |
|---|---|---|---|---|---|
| 2 | 404508.3.j | 468 | 548 | reverse 3 | TM |
| 2 | 404508.3.j | 549 | 626 | forward 3 | SP |
| 2 | 404508.3.j | 348 | 428 | reverse 3 | TM |
| 2 | 404508.3.j | 430 | 513 | reverse 1 | SP |
| 2 | 404508.3.j | 459 | 539 | forward 3 | TM |
| 3 | 441227.2.j | 523 | 606 | forward 1 | SP |
| 6 | 13039.2 | 1507 | 1581 | forward 1 | TM |
| 7 | 238005.4 | 887 | 973 | forward 2 | SP |
| 7 | 238005.4 | 1691 | 1783 | forward 2 | SP |
| 7 | 238005.4 | 1583 | 1678 | forward 2 | SP |
| 7 | 238005.4 | 566 | 643 | forward 2 | TM |
| 9 | 348094.6.j | 1994 | 2077 | forward 2 | TM |
| 9 | 348094.6.j | 1871 | 1948 | reverse 2 | TM |
| 9 | 348094.6.j | 2046 | 2117 | forward 3 | TM |
| 9 | 348094.6.j | 2112 | 2195 | reverse 3 | TM |
| 9 | 348094.6.j | 2035 | 2115 | forward 1 | TM |
| 9 | 348094.6.j | 2920 | 3000 | forward 1 | TM |
| 10 | 233678.1 | 2412 | 2495 | forward 3 | SP |
| 11 | 312243.1 | 528 | 614 | forward 3 | SP |
| 16 | 346078.5.j | 1406 | 1492 | forward 2 | TM |
| 17 | 394637.1.j | 36 | 122 | forward 3 | TM |
| 18 | 222429.3 | 646 | 726 | forward 1 | SP |
| 19 | 366739.2 | 791 | 880 | forward 2 | SP |
| 19 | 366739.2 | 1115 | 1207 | forward 2 | SP |
| 20 | 474635.6 | 1293 | 1370 | forward 3 | TM |
| 22 | 407090.5.j | 2955 | 3038 | reverse 3 | SP |
| 24 | 411449.2 | 283 | 366 | forward 1 | SP |
| 27 | 445433.2 | 123 | 203 | forward 3 | TM |
| 29 | 257121.2 | 3656 | 3736 | forward 2 | TM |
| 29 | 257121.2 | 851 | 934 | forward 2 | TM |
| 46 | 338992.1 | 1601 | 1690 | forward 2 | SP |
| 46 | 338992.1 | 1478 | 1561 | forward 2 | SP |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 4 | 277927.2 | 2706884H1 | 1 | 199 |
| 4 | 277927.2 | 4317170F6 | 1 | 384 |
| 4 | 277927.2 | 2706884F6 | 1 | 320 |
| 4 | 277927.2 | g1062337 | 36 | 388 |
| 4 | 277927.2 | g1064195 | 39 | 230 |
| 4 | 277927.2 | 2862653H1 | 42 | 117 |
| 4 | 277927.2 | 494376H1 | 53 | 305 |
| 4 | 277927.2 | 494376R7 | 53 | 383 |
| 4 | 277927.2 | 494376R6 | 53 | 492 |
| 4 | 277927.2 | g1275473 | 57 | 586 |
| 4 | 277927.2 | g2055898 | 59 | 543 |
| 4 | 277927.2 | 2707487H1 | 64 | 162 |
| 4 | 277927.2 | 491071H1 | 74 | 319 |
| 4 | 277927.2 | g1975146 | 80 | 403 |
| 4 | 277927.2 | g1142488 | 84 | 471 |
| 4 | 277927.2 | g1839772 | 84 | 214 |
| 4 | 277927.2 | g1012520 | 84 | 292 |
| 4 | 277927.2 | 3396290H1 | 86 | 344 |
| 4 | 277927.2 | 5896091H1 | 246 | 523 |
| 4 | 277927.2 | g946240 | 246 | 551 |
| 4 | 277927.2 | g943428 | 247 | 449 |
| 5 | 475311.1 | 1948746H1 | 1 | 233 |
| 5 | 475311.1 | g614088 | 12 | 271 |
| 5 | 475311.1 | 5644994H1 | 32 | 287 |
| 5 | 475311.1 | g645077 | 37 | 260 |
| 5 | 475311.1 | g668814 | 40 | 366 |
| 5 | 475311.1 | g645713 | 42 | 263 |
| 5 | 475311.1 | g645045 | 42 | 359 |
| 5 | 475311.1 | 4297187H1 | 46 | 315 |
| 5 | 475311.1 | 3486028H1 | 53 | 385 |
| 5 | 475311.1 | g683448 | 54 | 438 |
| 5 | 475311.1 | g645139 | 54 | 252 |
| 5 | 475311.1 | g645138 | 54 | 316 |
| 5 | 475311.1 | g670408 | 54 | 433 |
| 5 | 475311.1 | g646240 | 56 | 444 |
| 5 | 475311.1 | g562499 | 56 | 444 |
| 5 | 475311.1 | g4113162 | 56 | 511 |
| 5 | 475311.1 | g2003099 | 59 | 473 |
| 5 | 475311.1 | g815985 | 71 | 481 |
| 5 | 475311.1 | g674342 | 77 | 421 |
| 5 | 475311.1 | 4151149H1 | 88 | 355 |
| 5 | 475311.1 | 3843963H1 | 88 | 393 |
| 5 | 475311.1 | 5029373H1 | 92 | 346 |
| 5 | 475311.1 | 3584334H1 | 98 | 435 |
| 5 | 475311.1 | g3015753 | 105 | 441 |
| 5 | 475311.1 | g831378 | 105 | 457 |
| 5 | 475311.1 | g900504 | 105 | 401 |
| 5 | 475311.1 | g817737 | 105 | 402 |
| 5 | 475311.1 | g817003 | 105 | 418 |
| 5 | 475311.1 | g812268 | 105 | 397 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 5 | 475311.1 | g2052991 | 105 | 462 |
| 5 | 475311.1 | g3922575 | 105 | 509 |
| 5 | 475311.1 | g2884797 | 105 | 548 |
| 5 | 475311.1 | g2354824 | 105 | 375 |
| 5 | 475311.1 | g2788503 | 105 | 214 |
| 5 | 475311.1 | g797560 | 105 | 342 |
| 5 | 475311.1 | g820611 | 105 | 346 |
| 5 | 475311.1 | g821223 | 105 | 347 |
| 5 | 475311.1 | g1678742 | 138 | 543 |
| 5 | 475311.1 | 5590296H1 | 150 | 396 |
| 5 | 475311.1 | 2446939H1 | 155 | 407 |
| 5 | 475311.1 | 4178404H1 | 164 | 434 |
| 5 | 475311.1 | 4970343H1 | 179 | 455 |
| 5 | 475311.1 | g1720508 | 218 | 502 |
| 5 | 475311.1 | 4816085H1 | 242 | 326 |
| 5 | 475311.1 | g1296172 | 248 | 789 |
| 5 | 475311.1 | 5778021H1 | 256 | 544 |
| 5 | 475311.1 | 1579546F6 | 263 | 712 |
| 5 | 475311.1 | 1227469H1 | 263 | 535 |
| 5 | 475311.1 | 1951396H1 | 287 | 527 |
| 5 | 475311.1 | g1980791 | 307 | 631 |
| 5 | 475311.1 | 3369048H1 | 322 | 455 |
| 5 | 475311.1 | 261669H1 | 325 | 559 |
| 5 | 475311.1 | 1005556H1 | 344 | 651 |
| 5 | 475311.1 | g1987083 | 364 | 686 |
| 5 | 475311.1 | 3153327H1 | 364 | 683 |
| 5 | 475311.1 | 5086472H1 | 372 | 552 |
| 5 | 475311.1 | g2038507 | 365 | 782 |
| 5 | 475311.1 | 3791124H1 | 366 | 623 |
| 5 | 475311.1 | 3342191H1 | 383 | 665 |
| 5 | 475311.1 | 2727001H1 | 403 | 658 |
| 5 | 475311.1 | 5691335H1 | 421 | 697 |
| 5 | 475311.1 | 2246168H1 | 434 | 702 |
| 5 | 475311.1 | 4186012H1 | 582 | 933 |
| 5 | 475311.1 | 2630947H1 | 653 | 899 |
| 5 | 475311.1 | g1444007 | 793 | 1005 |
| 6 | 13039.2 | 4189083H1 | 3573 | 3883 |
| 6 | 13039.2 | g3802900 | 3582 | 3888 |
| 6 | 13039.2 | 2881380H1 | 3588 | 3897 |
| 6 | 13039.2 | g864431 | 3590 | 3852 |
| 6 | 13039.2 | g1481926 | 3594 | 3881 |
| 6 | 13039.2 | g667472 | 3600 | 3883 |
| 6 | 13039.2 | g3154327 | 3607 | 3891 |
| 6 | 13039.2 | 4583956H1 | 3611 | 3880 |
| 6 | 13039.2 | g776778 | 3615 | 3887 |
| 6 | 13039.2 | 2381092H1 | 3622 | 3854 |
| 6 | 13039.2 | 3810284H1 | 3623 | 3895 |
| 6 | 13039.2 | 4906362H2 | 3624 | 3904 |
| 6 | 13039.2 | 2202368H1 | 3632 | 3875 |
| 6 | 13039.2 | g4072852 | 3636 | 4070 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 2561132H1 | 3646 | 3748 |
| 6 | 13039.2 | 1294323F1 | 3648 | 3884 |
| 6 | 13039.2 | 1294323H1 | 3648 | 3780 |
| 6 | 13039.2 | 446234H1 | 3649 | 3871 |
| 6 | 13039.2 | 4774923H1 | 3650 | 3935 |
| 6 | 13039.2 | g789650 | 3651 | 3882 |
| 6 | 13039.2 | g1220073 | 3652 | 3888 |
| 6 | 13039.2 | g776622 | 3658 | 3886 |
| 6 | 13039.2 | g1522080 | 3663 | 3883 |
| 6 | 13039.2 | 449477H1 | 3667 | 3875 |
| 6 | 13039.2 | 2293709H1 | 3669 | 3884 |
| 6 | 13039.2 | g1114992 | 3695 | 3885 |
| 6 | 13039.2 | g667182 | 3700 | 3883 |
| 6 | 13039.2 | g1162269 | 3703 | 4063 |
| 6 | 13039.2 | g1773892 | 3703 | 4081 |
| 6 | 13039.2 | g3887238 | 3705 | 4083 |
| 6 | 13039.2 | 3542520H1 | 3707 | 3870 |
| 6 | 13039.2 | 4633040H1 | 3727 | 4003 |
| 6 | 13039.2 | g1128515 | 3763 | 4081 |
| 6 | 13039.2 | 2804481H1 | 3793 | 3899 |
| 6 | 13039.2 | 556467R6 | 3816 | 3871 |
| 6 | 13039.2 | 556467H1 | 3816 | 3871 |
| 6 | 13039.2 | g683155 | 3819 | 4075 |
| 6 | 13039.2 | g878104 | 3824 | 4075 |
| 6 | 13039.2 | 2945643H1 | 3841 | 4076 |
| 6 | 13039.2 | g821501 | 3869 | 4087 |
| 6 | 13039.2 | g782508 | 3869 | 4083 |
| 6 | 13039.2 | 2590326H2 | 4016 | 4075 |
| 6 | 13039.2 | 453597IT6 | 1081 | 1596 |
| 6 | 13039.2 | 2887511H1 | 1137 | 1386 |
| 6 | 13039.2 | 5060591H1 | 1078 | 1278 |
| 6 | 13039.2 | 2261086H1 | 1154 | 1385 |
| 6 | 13039.2 | 3535673H1 | 1214 | 1493 |
| 6 | 13039.2 | 1951291H1 | 1264 | 1453 |
| 6 | 13039.2 | 488798H1 | 1330 | 1587 |
| 6 | 13039.2 | 2634402H1 | 1381 | 1594 |
| 6 | 13039.2 | 2135368H1 | 1421 | 1684 |
| 6 | 13039.2 | 2135368F6 | 1421 | 1863 |
| 6 | 13039.2 | 2837860T6 | 3920 | 4029 |
| 6 | 13039.2 | 2837860F6 | 3927 | 4069 |
| 6 | 13039.2 | 2837860H1 | 3927 | 4069 |
| 6 | 13039.2 | 3536266H1 | 3936 | 4027 |
| 6 | 13039.2 | 3536268H1 | 3937 | 4038 |
| 6 | 13039.2 | g776777 | 3055 | 3279 |
| 6 | 13039.2 | g1856268 | 3100 | 3505 |
| 6 | 13039.2 | 2604507H1 | 3110 | 3384 |
| 6 | 13039.2 | 5065865H1 | 3071 | 3344 |
| 6 | 13039.2 | 3472443H1 | 3077 | 3330 |
| 6 | 13039.2 | 5017370H1 | 3114 | 3384 |
| 6 | 13039.2 | 2767212H1 | 3090 | 3334 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 790129R1 | 3121 | 3597 |
| 6 | 13039.2 | 2690375H1 | 3094 | 3357 |
| 6 | 13039.2 | 2411049H1 | 3099 | 3334 |
| 6 | 13039.2 | g1976982 | 1594 | 1956 |
| 6 | 13039.2 | 2642416H1 | 1599 | 1728 |
| 6 | 13039.2 | g1476955 | 1606 | 2005 |
| 6 | 13039.2 | g1062967 | 1606 | 1901 |
| 6 | 13039.2 | g1481925 | 1606 | 1784 |
| 6 | 13039.2 | 3136852H1 | 1738 | 2037 |
| 6 | 13039.2 | 2500241H1 | 1778 | 2029 |
| 6 | 13039.2 | 2788782H1 | 1863 | 2121 |
| 6 | 13039.2 | 874642H1 | 1862 | 2024 |
| 6 | 13039.2 | g1200951 | 1869 | 2062 |
| 6 | 13039.2 | 4535987H1 | 1879 | 2008 |
| 6 | 13039.2 | 4536041H1 | 1884 | 2151 |
| 6 | 13039.2 | 1921318T6 | 3558 | 4028 |
| 6 | 13039.2 | g2458968 | 3562 | 3880 |
| 6 | 13039.2 | 3434402H1 | 1 | 237 |
| 6 | 13039.2 | g1774870 | 25 | 371 |
| 6 | 13039.2 | 5090377H1 | 28 | 270 |
| 6 | 13039.2 | 5090377F6 | 28 | 341 |
| 6 | 13039.2 | g2243533 | 71 | 456 |
| 6 | 13039.2 | 4535971H1 | 377 | 641 |
| 6 | 13039.2 | 4535971F6 | 377 | 900 |
| 6 | 13039.2 | 2738090H1 | 634 | 871 |
| 6 | 13039.2 | 2738090F6 | 634 | 1152 |
| 6 | 13039.2 | 4032617H1 | 652 | 788 |
| 6 | 13039.2 | 3247382H1 | 687 | 986 |
| 6 | 13039.2 | 2886124H1 | 2974 | 3263 |
| 6 | 13039.2 | 2886134H1 | 2974 | 3255 |
| 6 | 13039.2 | 2882975H1 | 2974 | 3267 |
| 6 | 13039.2 | 731806R1 | 2984 | 3499 |
| 6 | 13039.2 | 3715202H1 | 2984 | 3092 |
| 6 | 13039.2 | g1993687 | 2985 | 3312 |
| 6 | 13039.2 | 731806H1 | 2984 | 3258 |
| 6 | 13039.2 | 2468144H1 | 2990 | 3229 |
| 6 | 13039.2 | 1907815H1 | 2991 | 3218 |
| 6 | 13039.2 | 4148570H1 | 3006 | 3269 |
| 6 | 13039.2 | g778804 | 3015 | 3330 |
| 6 | 13039.2 | 3256396H1 | 3017 | 3269 |
| 6 | 13039.2 | g789649 | 3021 | 3255 |
| 6 | 13039.2 | 4983180H1 | 3028 | 3311 |
| 6 | 13039.2 | g865926 | 3030 | 3351 |
| 6 | 13039.2 | 4983190H1 | 3028 | 3302 |
| 6 | 13039.2 | 998221R1 | 3039 | 3594 |
| 6 | 13039.2 | 998221H1 | 3039 | 3328 |
| 6 | 13039.2 | 4650828H1 | 3041 | 3328 |
| 6 | 13039.2 | g776667 | 3055 | 3330 |
| 6 | 13039.2 | 3699551H1 | 759 | 1043 |
| 6 | 13039.2 | 2784137H1 | 768 | 999 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 5092174H1 | 863 | 1136 |
| 6 | 13039.2 | 2736123H1 | 901 | 1143 |
| 6 | 13039.2 | 2736123F6 | 901 | 1374 |
| 6 | 13039.2 | 2841123H2 | 964 | 1052 |
| 6 | 13039.2 | g1801914 | 984 | 1330 |
| 6 | 13039.2 | 3627415H1 | 990 | 1242 |
| 6 | 13039.2 | 2049675H1 | 1064 | 1318 |
| 6 | 13039.2 | 2783785F6 | 1067 | 1509 |
| 6 | 13039.2 | 2783785H1 | 1067 | 1319 |
| 6 | 13039.2 | 4764557H1 | 2700 | 2895 |
| 6 | 13039.2 | g862228 | 2711 | 2935 |
| 6 | 13039.2 | 4977405H1 | 2718 | 2987 |
| 6 | 13039.2 | 4324144H1 | 2725 | 3007 |
| 6 | 13039.2 | 5195156H1 | 2730 | 2831 |
| 6 | 13039.2 | 4625438H1 | 2738 | 2992 |
| 6 | 13039.2 | 2881531H1 | 2740 | 3023 |
| 6 | 13039.2 | 1545834H1 | 2740 | 2953 |
| 6 | 13039.2 | 3214807H1 | 2740 | 2990 |
| 6 | 13039.2 | 1312356H1 | 2750 | 2954 |
| 6 | 13039.2 | 1312307H1 | 2750 | 2979 |
| 6 | 13039.2 | g1238200 | 3569 | 3881 |
| 6 | 13039.2 | g792128 | 3570 | 3881 |
| 6 | 13039.2 | g3109745 | 3571 | 3878 |
| 6 | 13039.2 | g789837 | 2241 | 2520 |
| 6 | 13039.2 | g666844 | 2248 | 2506 |
| 6 | 13039.2 | g793332 | 2260 | 2481 |
| 6 | 13039.2 | g862227 | 2260 | 2522 |
| 6 | 13039.2 | 3321110H1 | 2293 | 2570 |
| 6 | 13039.2 | g1636044 | 2306 | 2530 |
| 6 | 13039.2 | 3738558H1 | 2316 | 2611 |
| 6 | 13039.2 | 5700029H1 | 2317 | 2575 |
| 6 | 13039.2 | 5030623H1 | 2361 | 2636 |
| 6 | 13039.2 | 3321569H2 | 2364 | 2614 |
| 6 | 13039.2 | 3484339H1 | 2365 | 2701 |
| 6 | 13039.2 | g317338 | 2375 | 2655 |
| 6 | 13039.2 | 4922854H1 | 2377 | 2683 |
| 6 | 13039.2 | 1382370H1 | 2377 | 2615 |
| 6 | 13039.2 | 3364673H1 | 2393 | 2653 |
| 6 | 13039.2 | 4014451H1 | 2393 | 2645 |
| 6 | 13039.2 | 4296168H1 | 2429 | 2676 |
| 6 | 13039.2 | 4296144H1 | 2429 | 2683 |
| 6 | 13039.2 | 4295505H1 | 2429 | 2696 |
| 6 | 13039.2 | 4775793H1 | 2466 | 2741 |
| 6 | 13039.2 | 5190065H1 | 1477 | 1620 |
| 6 | 13039.2 | g2013579 | 1518 | 1807 |
| 6 | 13039.2 | 3201887H1 | 1570 | 1835 |
| 6 | 13039.2 | g708914 | 1436 | 1737 |
| 6 | 13039.2 | g691657 | 1479 | 1801 |
| 6 | 13039.2 | 489678H1 | 1579 | 1818 |
| 6 | 13039.2 | 044642H1 | 2172 | 2328 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 3129828H1 | 2163 | 2472 |
| 6 | 13039.2 | 5702193H1 | 2183 | 2423 |
| 6 | 13039.2 | g866587 | 2185 | 2551 |
| 6 | 13039.2 | 4968731H1 | 2217 | 2474 |
| 6 | 13039.2 | 4969450H1 | 2217 | 2498 |
| 6 | 13039.2 | g434208 | 2221 | 2510 |
| 6 | 13039.2 | g666222 | 3533 | 3878 |
| 6 | 13039.2 | 1861085T6 | 3554 | 3841 |
| 6 | 13039.2 | g1476956 | 3554 | 3883 |
| 6 | 13039.2 | 2472134H1 | 1917 | 2158 |
| 6 | 13039.2 | 2460193H1 | 1973 | 2223 |
| 6 | 13039.2 | 5544753H1 | 1977 | 2125 |
| 6 | 13039.2 | 1894461H1 | 1994 | 2224 |
| 6 | 13039.2 | 4989013H1 | 2004 | 2267 |
| 6 | 13039.2 | 2504842H1 | 2018 | 2250 |
| 6 | 13039.2 | 3817686H1 | 2039 | 2289 |
| 6 | 13039.2 | 1709583F6 | 2086 | 2637 |
| 6 | 13039.2 | 1709583H1 | 2086 | 2292 |
| 6 | 13039.2 | 5496588H1 | 2117 | 2221 |
| 6 | 13039.2 | g618254 | 2133 | 2456 |
| 6 | 13039.2 | 2302357H2 | 2134 | 2358 |
| 6 | 13039.2 | 3076667H1 | 2142 | 2408 |
| 6 | 13039.2 | g4096030 | 3440 | 3883 |
| 6 | 13039.2 | g4394182 | 3444 | 3889 |
| 6 | 13039.2 | g3298631 | 3450 | 3883 |
| 6 | 13039.2 | g4152977 | 3452 | 3888 |
| 6 | 13039.2 | 4150667H1 | 3456 | 3752 |
| 6 | 13039.2 | 5036313H1 | 3454 | 3744 |
| 6 | 13039.2 | g4110089 | 3455 | 3887 |
| 6 | 13039.2 | g3110528 | 3467 | 3881 |
| 6 | 13039.2 | g3884368 | 3468 | 3883 |
| 6 | 13039.2 | g2631703 | 3472 | 3882 |
| 6 | 13039.2 | g4329108 | 3483 | 3871 |
| 6 | 13039.2 | g4307404 | 3484 | 3891 |
| 6 | 13039.2 | g3887724 | 3487 | 3890 |
| 6 | 13039.2 | 5679544H1 | 3488 | 3784 |
| 6 | 13039.2 | 4643394H1 | 3488 | 3741 |
| 6 | 13039.2 | 231550R1 | 3496 | 3891 |
| 6 | 13039.2 | 231550H1 | 3496 | 3719 |
| 6 | 13039.2 | g2659034 | 3504 | 3678 |
| 6 | 13039.2 | 3635404H1 | 3504 | 3814 |
| 6 | 13039.2 | 2736123T6 | 3515 | 3839 |
| 6 | 13039.2 | 231550F1 | 3523 | 3882 |
| 6 | 13039.2 | g3596402 | 3524 | 3881 |
| 6 | 13039.2 | 2736217H1 | 3526 | 3796 |
| 6 | 13039.2 | 2368391H2 | 3323 | 3556 |
| 6 | 13039.2 | 4761918H1 | 3324 | 3582 |
| 6 | 13039.2 | 3162730H1 | 3331 | 3637 |
| 6 | 13039.2 | 2911248H1 | 3338 | 3625 |
| 6 | 13039.2 | 780353H1 | 3345 | 3665 |

TABLE 4

| SEQ ID NO: | T mplate ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 1233680H1 | 3349 | 3681 |
| 6 | 13039.2 | 1233680F1 | 3349 | 3971 |
| 6 | 13039.2 | 3792837H1 | 3351 | 3635 |
| 6 | 13039.2 | g2357493 | 3351 | 3602 |
| 6 | 13039.2 | 2005695H1 | 3354 | 3432 |
| 6 | 13039.2 | g3087116 | 3355 | 3891 |
| 6 | 13039.2 | 3038679H1 | 3359 | 3631 |
| 6 | 13039.2 | 731806F1 | 3361 | 3882 |
| 6 | 13039.2 | g3070256 | 3362 | 3878 |
| 6 | 13039.2 | g3049253 | 3364 | 3885 |
| 6 | 13039.2 | g3049251 | 3366 | 3885 |
| 6 | 13039.2 | g3070255 | 3369 | 3878 |
| 6 | 13039.2 | g3076969 | 3370 | 3878 |
| 6 | 13039.2 | g4292218 | 3374 | 3875 |
| 6 | 13039.2 | 2379289H1 | 3376 | 3605 |
| 6 | 13039.2 | 6026190H1 | 3375 | 3685 |
| 6 | 13039.2 | 2887871H1 | 3378 | 3685 |
| 6 | 13039.2 | g3841328 | 3384 | 3878 |
| 6 | 13039.2 | g4073143 | 3385 | 3881 |
| 6 | 13039.2 | 602339H1 | 3386 | 3668 |
| 6 | 13039.2 | g2901329 | 3390 | 3885 |
| 6 | 13039.2 | g4153184 | 3401 | 3882 |
| 6 | 13039.2 | 1862447T6 | 3417 | 3836 |
| 6 | 13039.2 | 2738090T6 | 3429 | 3840 |
| 6 | 13039.2 | 1709583T6 | 3431 | 3837 |
| 6 | 13039.2 | 2870017H1 | 3229 | 3528 |
| 6 | 13039.2 | 2859049H1 | 3229 | 3512 |
| 6 | 13039.2 | 1822377H1 | 3237 | 3499 |
| 6 | 13039.2 | 3601353H1 | 3245 | 3577 |
| 6 | 13039.2 | 2760117H1 | 3247 | 3543 |
| 6 | 13039.2 | 2751935H1 | 3247 | 3520 |
| 6 | 13039.2 | 1921318H1 | 3255 | 3535 |
| 6 | 13039.2 | 1921318R6 | 3255 | 3696 |
| 6 | 13039.2 | 4797434H1 | 3262 | 3539 |
| 6 | 13039.2 | g778894 | 3266 | 3611 |
| 6 | 13039.2 | g855990 | 3266 | 3591 |
| 6 | 13039.2 | 2135368T6 | 3266 | 3840 |
| 6 | 13039.2 | 3975369H1 | 3275 | 3387 |
| 6 | 13039.2 | 3442784H1 | 3283 | 3547 |
| 6 | 13039.2 | 1832656H1 | 3299 | 3596 |
| 6 | 13039.2 | 2219316H1 | 3317 | 3573 |
| 6 | 13039.2 | 4761911H1 | 3323 | 3598 |
| 6 | 13039.2 | g1238488 | 3188 | 3379 |
| 6 | 13039.2 | g877900 | 3195 | 3513 |
| 6 | 13039.2 | g2148718 | 3196 | 3847 |
| 6 | 13039.2 | 1736290H1 | 3199 | 3439 |
| 6 | 13039.2 | 1734801H1 | 3199 | 3448 |
| 6 | 13039.2 | 1734817H1 | 3199 | 3447 |
| 6 | 13039.2 | 3812780H1 | 3201 | 3506 |
| 6 | 13039.2 | 3814117H1 | 3202 | 3367 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | 2404648H1 | 3213 | 3450 |
| 6 | 13039.2 | 1736463H1 | 3215 | 3509 |
| 6 | 13039.2 | 5019558H1 | 3217 | 3470 |
| 6 | 13039.2 | 5559285H1 | 3217 | 3498 |
| 6 | 13039.2 | 793056H1 | 3123 | 3357 |
| 6 | 13039.2 | 790129H1 | 3134 | 3376 |
| 6 | 13039.2 | 2671569H1 | 3169 | 3272 |
| 6 | 13039.2 | 4255130H1 | 3176 | 3267 |
| 6 | 13039.2 | 4164759H1 | 3170 | 3434 |
| 6 | 13039.2 | 4255178H1 | 3175 | 3459 |
| 6 | 13039.2 | 1861085F6 | 3562 | 3882 |
| 6 | 13039.2 | g1203064 | 3563 | 3884 |
| 6 | 13039.2 | 1861177H1 | 3563 | 3888 |
| 6 | 13039.2 | 2803220H1 | 2525 | 2792 |
| 6 | 13039.2 | 4879807H1 | 2476 | 2740 |
| 6 | 13039.2 | 1627449H1 | 2527 | 2661 |
| 6 | 13039.2 | 2660611H1 | 2481 | 2734 |
| 6 | 13039.2 | 3449673H1 | 2502 | 2757 |
| 6 | 13039.2 | 5039738H2 | 2535 | 2749 |
| 6 | 13039.2 | 4595615H1 | 2538 | 2813 |
| 6 | 13039.2 | 1862447F6 | 2566 | 3056 |
| 6 | 13039.2 | 1862447H1 | 2566 | 2831 |
| 6 | 13039.2 | 4251362H1 | 2587 | 2852 |
| 6 | 13039.2 | g864534 | 2601 | 2918 |
| 6 | 13039.2 | g1949105 | 2608 | 2883 |
| 6 | 13039.2 | 4831252H1 | 2612 | 2740 |
| 6 | 13039.2 | 4830165H1 | 2612 | 2824 |
| 6 | 13039.2 | g1522198 | 2629 | 2917 |
| 6 | 13039.2 | 2962002H1 | 2642 | 2938 |
| 6 | 13039.2 | 4111111H1 | 2668 | 2941 |
| 6 | 13039.2 | 4581818H1 | 2674 | 2970 |
| 6 | 13039.2 | 3480168H1 | 2674 | 2966 |
| 6 | 13039.2 | 059851H1 | 2688 | 2875 |
| 6 | 13039.2 | 3463785H1 | 2827 | 3087 |
| 6 | 13039.2 | 1379162H1 | 2833 | 3074 |
| 6 | 13039.2 | 827278H1 | 2833 | 3141 |
| 6 | 13039.2 | 1379162F1 | 2833 | 3380 |
| 6 | 13039.2 | 827278R1 | 2833 | 3428 |
| 6 | 13039.2 | 4970664H1 | 2852 | 3138 |
| 6 | 13039.2 | 3765903H1 | 2856 | 3173 |
| 6 | 13039.2 | g828803 | 2876 | 3107 |
| 6 | 13039.2 | 2912660H1 | 2876 | 3139 |
| 6 | 13039.2 | 2945011H1 | 2929 | 3231 |
| 6 | 13039.2 | 1900343H1 | 2930 | 3184 |
| 6 | 13039.2 | 4177622H1 | 2939 | 3215 |
| 6 | 13039.2 | 2714979H1 | 2941 | 3205 |
| 6 | 13039.2 | 5098921H1 | 2946 | 3241 |
| 6 | 13039.2 | 4589878H1 | 2963 | 3194 |
| 6 | 13039.2 | 1350496H1 | 2966 | 3227 |
| 6 | 13039.2 | 1350496F1 | 2966 | 3639 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 6 | 13039.2 | g793050 | 2751 | 2936 |
| 6 | 13039.2 | 2396484H1 | 2751 | 2985 |
| 6 | 13039.2 | 3677377H1 | 2764 | 2931 |
| 6 | 13039.2 | 1613610H1 | 2765 | 2976 |
| 6 | 13039.2 | 3677367H1 | 2765 | 3072 |
| 6 | 13039.2 | 3813522H1 | 2772 | 3037 |
| 6 | 13039.2 | 4688970H1 | 2782 | 3032 |
| 6 | 13039.2 | 4889047H1 | 2811 | 3096 |
| 6 | 13039.2 | 4550845H1 | 2811 | 2992 |
| 6 | 13039.2 | g1986776 | 2811 | 3068 |
| 6 | 13039.2 | 2943108H1 | 2816 | 3136 |
| 6 | 13039.2 | 919863H1 | 2816 | 3094 |
| 6 | 13039.2 | 2220586H1 | 2818 | 3075 |
| 7 | 238005.4 | 3494842H1 | 1 | 146 |
| 7 | 238005.4 | 1513955H1 | 19 | 211 |
| 7 | 238005.4 | 1513931H1 | 19 | 189 |
| 7 | 238005.4 | 1895354H1 | 24 | 255 |
| 7 | 238005.4 | 3216547H1 | 53 | 287 |
| 7 | 238005.4 | 2492948H1 | 210 | 525 |
| 7 | 238005.4 | 3393775H1 | 226 | 487 |
| 7 | 238005.4 | 3110735H1 | 237 | 520 |
| 7 | 238005.4 | 2114053H1 | 238 | 500 |
| 7 | 238005.4 | 2639154H1 | 229 | 480 |
| 7 | 238005.4 | 2639154F6 | 229 | 615 |
| 7 | 238005.4 | 3489530H1 | 251 | 540 |
| 7 | 238005.4 | 2452873H1 | 253 | 499 |
| 7 | 238005.4 | 2452873F6 | 253 | 419 |
| 7 | 238005.4 | 3149681H1 | 258 | 499 |
| 7 | 238005.4 | 2952379H1 | 265 | 542 |
| 7 | 238005.4 | 2466034H1 | 291 | 507 |
| 7 | 238005.4 | 3074678H1 | 330 | 611 |
| 7 | 238005.4 | g2003143 | 378 | 830 |
| 7 | 238005.4 | 2310518R6 | 454 | 935 |
| 7 | 238005.4 | 2310518H1 | 454 | 735 |
| 7 | 238005.4 | 2466818H1 | 500 | 732 |
| 7 | 238005.4 | 3243701H1 | 503 | 764 |
| 7 | 238005.4 | 3108380H1 | 506 | 773 |
| 7 | 238005.4 | 4250192H1 | 506 | 710 |
| 7 | 238005.4 | 3688838H1 | 527 | 810 |
| 7 | 238005.4 | 2312916H1 | 543 | 816 |
| 7 | 238005.4 | 4087821H1 | 570 | 878 |
| 7 | 238005.4 | g3173776 | 593 | 909 |
| 7 | 238005.4 | 3960672H2 | 599 | 874 |
| 7 | 238005.4 | 1830753H1 | 644 | 897 |
| 7 | 238005.4 | 3936919H1 | 690 | 871 |
| 7 | 238005.4 | 3183381H1 | 691 | 942 |
| 7 | 238005.4 | 3165765H1 | 714 | 1037 |
| 7 | 238005.4 | 3231218H1 | 762 | 1012 |
| 7 | 238005.4 | 5273755H1 | 763 | 1032 |
| 7 | 238005.4 | 4710712H1 | 876 | 997 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 7 | 238005.4 | g847384 | 989 | 1313 |
| 7 | 238005.4 | 5810260H1 | 999 | 1303 |
| 7 | 238005.4 | 2046234F6 | 1017 | 1417 |
| 7 | 238005.4 | 4547319H1 | 1024 | 1320 |
| 7 | 238005.4 | 5302724H1 | 1041 | 1284 |
| 7 | 238005.4 | 5267284H1 | 1053 | 1283 |
| 7 | 238005.4 | 4862949H1 | 1072 | 1245 |
| 7 | 238005.4 | 2893053H1 | 1078 | 1355 |
| 7 | 238005.4 | 2893053F6 | 1078 | 1449 |
| 7 | 238005.4 | 3408366H1 | 1120 | 1378 |
| 7 | 238005.4 | 5427753H1 | 1132 | 1350 |
| 7 | 238005.4 | 2046234H1 | 1135 | 1417 |
| 7 | 238005.4 | g846587 | 1153 | 1408 |
| 7 | 238005.4 | 1978835H1 | 1173 | 1431 |
| 7 | 238005.4 | 5187046H1 | 1212 | 1303 |
| 7 | 238005.4 | 5049433F6 | 1212 | 1740 |
| 7 | 238005.4 | 5069619H1 | 1213 | 1457 |
| 7 | 238005.4 | 079996H1 | 1221 | 1447 |
| 7 | 238005.4 | 2755107H1 | 1244 | 1546 |
| 7 | 238005.4 | 2994580H1 | 1250 | 1546 |
| 7 | 238005.4 | 1357968H1 | 1253 | 1465 |
| 7 | 238005.4 | 2486058H1 | 1290 | 1528 |
| 7 | 238005.4 | 3155023H1 | 1308 | 1607 |
| 7 | 238005.4 | 1405122H1 | 1358 | 1634 |
| 7 | 238005.4 | 074212H1 | 1360 | 1547 |
| 7 | 238005.4 | g1775652 | 1362 | 1565 |
| 7 | 238005.4 | 5426560H1 | 1368 | 1643 |
| 7 | 238005.4 | 2452873T6 | 1376 | 1902 |
| 7 | 238005.4 | 5597619H1 | 1385 | 1634 |
| 7 | 238005.4 | 2640268T6 | 1394 | 1890 |
| 7 | 238005.4 | 2971688H1 | 1399 | 1715 |
| 7 | 238005.4 | 2310518T6 | 1425 | 1902 |
| 7 | 238005.4 | 522334H1 | 1421 | 1673 |
| 7 | 238005.4 | 4787913H1 | 1436 | 1716 |
| 7 | 238005.4 | 6093903H1 | 1439 | 1775 |
| 7 | 238005.4 | 3716621H1 | 1449 | 1751 |
| 7 | 238005.4 | 3435261H1 | 1451 | 1718 |
| 7 | 238005.4 | 3256708H1 | 1454 | 1740 |
| 7 | 238005.4 | 769094H1 | 1454 | 1685 |
| 7 | 238005.4 | g2526289 | 1458 | 1946 |
| 7 | 238005.4 | 4974673H1 | 1461 | 1728 |
| 7 | 238005.4 | 3870931H1 | 1471 | 1787 |
| 7 | 238005.4 | 4701192H1 | 1471 | 1746 |
| 7 | 238005.4 | g4113880 | 1481 | 1936 |
| 7 | 238005.4 | 5703513H1 | 1508 | 1795 |
| 7 | 238005.4 | 5867713H1 | 1510 | 1740 |
| 7 | 238005.4 | g3430405 | 1520 | 1936 |
| 7 | 238005.4 | g4304488 | 1527 | 1925 |
| 7 | 238005.4 | g3446639 | 1532 | 1952 |
| 7 | 238005.4 | g3843850 | 1534 | 1956 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 7 | 238005.4 | 5049433T6 | 1538 | 1924 |
| 7 | 238005.4 | g3895204 | 1538 | 1945 |
| 7 | 238005.4 | 3343709T6 | 1541 | 1905 |
| 7 | 238005.4 | g4267940 | 1548 | 1946 |
| 7 | 238005.4 | 2269156H1 | 1556 | 1832 |
| 7 | 238005.4 | 3211050H1 | 1556 | 1795 |
| 7 | 238005.4 | 3858314H1 | 1561 | 1865 |
| 7 | 238005.4 | 3858359H1 | 1561 | 1845 |
| 7 | 238005.4 | 3861114H1 | 1561 | 1868 |
| 7 | 238005.4 | 4215258H1 | 1567 | 1843 |
| 7 | 238005.4 | g1647982 | 1573 | 1944 |
| 7 | 238005.4 | 6106957H1 | 1583 | 1903 |
| 7 | 238005.4 | 4974030H1 | 1586 | 1853 |
| 7 | 238005.4 | g518289 | 1588 | 1946 |
| 7 | 238005.4 | g2526290 | 1592 | 1945 |
| 7 | 238005.4 | 4562130H1 | 1603 | 1871 |
| 7 | 238005.4 | 6093354H1 | 1607 | 1891 |
| 7 | 238005.4 | 2611577H1 | 1612 | 1852 |
| 7 | 238005.4 | g2003142 | 1617 | 1950 |
| 7 | 238005.4 | g2213309 | 1628 | 1935 |
| 7 | 238005.4 | g1893696 | 1633 | 1950 |
| 7 | 238005.4 | 2759732H1 | 1636 | 1924 |
| 7 | 238005.4 | g4243121 | 1640 | 1944 |
| 7 | 238005.4 | g4108171 | 1645 | 1945 |
| 7 | 238005.4 | g3765581 | 1650 | 1946 |
| 7 | 238005.4 | 3027260T6 | 1677 | 1899 |
| 7 | 238005.4 | 1213283H1 | 1736 | 1875 |
| 7 | 238005.4 | g846910 | 1739 | 1936 |
| 7 | 238005.4 | 2753427H1 | 1746 | 1944 |
| 7 | 238005.4 | 2150089H1 | 1756 | 1944 |
| 7 | 238005.4 | g1775754 | 1824 | 1936 |
| 7 | 238005.4 | 5288457H1 | 1860 | 1978 |
| 7 | 238005.4 | g4285357 | 1863 | 1946 |
| 10 | 233678.1 | 1803848F6 | 1839 | 2351 |
| 10 | 233678.1 | 2355751H1 | 1855 | 1998 |
| 10 | 233678.1 | 1923122H1 | 1895 | 2186 |
| 10 | 233678.1 | 1858559H1 | 1896 | 2188 |
| 10 | 233678.1 | 2949425H1 | 1896 | 2193 |
| 10 | 233678.1 | 1858559F6 | 1896 | 2326 |
| 10 | 233678.1 | 4542183H1 | 1896 | 2166 |
| 10 | 233678.1 | 3786222H1 | 1898 | 2215 |
| 10 | 233678.1 | 6091924H1 | 1943 | 2212 |
| 10 | 233678.1 | 4584888H1 | 1958 | 2251 |
| 10 | 233678.1 | 1651475H1 | 1991 | 2242 |
| 10 | 233678.1 | 3967280H1 | 2006 | 2242 |
| 10 | 233678.1 | 5623384H1 | 2022 | 2351 |
| 10 | 233678.1 | 5623284H1 | 2022 | 2314 |
| 10 | 233678.1 | 4948649H1 | 2031 | 2325 |
| 10 | 233678.1 | 1521957H1 | 2030 | 2222 |
| 10 | 233678.1 | 1726538H1 | 2034 | 2236 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 10 | 233678.1 | 1723946H1 | 2034 | 2252 |
| 10 | 233678.1 | 1605814H1 | 2037 | 2174 |
| 10 | 233678.1 | 5554120H1 | 2037 | 2321 |
| 10 | 233678.1 | 1874743F6 | 2056 | 2410 |
| 10 | 233678.1 | 1874743H1 | 2057 | 2352 |
| 10 | 233678.1 | 4857286H1 | 2097 | 2352 |
| 10 | 233678.1 | 1509348H1 | 2102 | 2301 |
| 10 | 233678.1 | 2242767H1 | 2114 | 2370 |
| 10 | 233678.1 | 2848660H1 | 2126 | 2428 |
| 10 | 233678.1 | 2686671H1 | 2129 | 2380 |
| 10 | 233678.1 | 1576284H1 | 2142 | 2375 |
| 10 | 233678.1 | 887052R1 | 2142 | 2769 |
| 10 | 233678.1 | 887052H1 | 2142 | 2447 |
| 10 | 233678.1 | 3785578H1 | 2149 | 2345 |
| 10 | 233678.1 | 3111926H1 | 2149 | 2473 |
| 10 | 233678.1 | 766862H1 | 2158 | 2406 |
| 10 | 233678.1 | 2227686H1 | 2171 | 2407 |
| 10 | 233678.1 | 2013238H1 | 2171 | 2277 |
| 10 | 233678.1 | 823920R1 | 2176 | 2721 |
| 10 | 233678.1 | 1583169H1 | 2176 | 2393 |
| 10 | 233678.1 | 823920H1 | 2176 | 2328 |
| 10 | 233678.1 | 1583137H1 | 2176 | 2390 |
| 10 | 233678.1 | g1426492 | 2182 | 2682 |
| 10 | 233678.1 | 2275987H1 | 2225 | 2478 |
| 10 | 233678.1 | 4730301H1 | 2231 | 2455 |
| 10 | 233678.1 | 176909H1 | 2231 | 2495 |
| 10 | 233678.1 | 1702710H1 | 2230 | 2441 |
| 10 | 233678.1 | 2845122H1 | 2231 | 2466 |
| 10 | 233678.1 | 5895659H1 | 2245 | 2544 |
| 10 | 233678.1 | 4703091H1 | 2244 | 2507 |
| 10 | 233678.1 | 4214988H1 | 2256 | 2569 |
| 10 | 233678.1 | 1622006H1 | 2271 | 2508 |
| 10 | 233678.1 | 1250975T6 | 2285 | 2941 |
| 10 | 233678.1 | 5874967H1 | 2299 | 2566 |
| 10 | 233678.1 | 3291543H1 | 2301 | 2556 |
| 10 | 233678.1 | 5874909H1 | 2300 | 2556 |
| 10 | 233678.1 | 3291543F6 | 2301 | 2749 |
| 10 | 233678.1 | 4588984H1 | 2318 | 2524 |
| 10 | 233678.1 | 4564713H1 | 2321 | 2549 |
| 10 | 233678.1 | 1607223F6 | 2329 | 2631 |
| 10 | 233678.1 | 1607223H1 | 2329 | 2561 |
| 10 | 233678.1 | 777428R1 | 2329 | 2946 |
| 10 | 233678.1 | 4196029H1 | 2328 | 2638 |
| 10 | 233678.1 | 777428H1 | 2329 | 2573 |
| 10 | 233678.1 | 3098571T6 | 2334 | 2942 |
| 10 | 233678.1 | 2008649H1 | 2355 | 2571 |
| 10 | 233678.1 | 1607223T6 | 2360 | 2927 |
| 10 | 233678.1 | 1803848T6 | 2373 | 2929 |
| 10 | 233678.1 | 2210370T6 | 2375 | 2928 |
| 10 | 233678.1 | 1701603H1 | 2387 | 2602 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 10 | 233678.1 | 5286984H1 | 2406 | 2602 |
| 10 | 233678.1 | 2355751T6 | 2414 | 2928 |
| 10 | 233678.1 | 285554H1 | 2419 | 2806 |
| 10 | 233678.1 | 285554R6 | 2420 | 2891 |
| 10 | 233678.1 | 3289542T6 | 2435 | 2926 |
| 10 | 233678.1 | 2518190T6 | 2436 | 2916 |
| 10 | 233678.1 | 285554T6 | 2452 | 2926 |
| 10 | 233678.1 | 4949921H1 | 2453 | 2734 |
| 10 | 233678.1 | 2814466T6 | 2451 | 2932 |
| 10 | 233678.1 | 3291543T6 | 2464 | 2930 |
| 10 | 233678.1 | 4197595H1 | 2466 | 2785 |
| 10 | 233678.1 | 5105034H1 | 2468 | 2740 |
| 10 | 233678.1 | g2963989 | 2472 | 2972 |
| 10 | 233678.1 | g2063675 | 2474 | 2971 |
| 10 | 233678.1 | g517590 | 2474 | 2968 |
| 10 | 233678.1 | 4950109H1 | 2484 | 2768 |
| 10 | 233678.1 | 1874743T6 | 2498 | 2928 |
| 10 | 233678.1 | g4107810 | 2500 | 2968 |
| 10 | 233678.1 | g4175509 | 2507 | 2977 |
| 10 | 233678.1 | g3411743 | 2509 | 2973 |
| 10 | 233678.1 | 1858559T6 | 2509 | 2930 |
| 10 | 233678.1 | 5848668H1 | 2517 | 2804 |
| 10 | 233678.1 | g4269824 | 2519 | 2968 |
| 10 | 233678.1 | 3578266T6 | 2525 | 2934 |
| 10 | 233678.1 | 2623196H1 | 2526 | 2783 |
| 10 | 233678.1 | g3770184 | 2541 | 2971 |
| 10 | 233678.1 | g3923313 | 2548 | 2973 |
| 10 | 233678.1 | 3274147T6 | 2551 | 2910 |
| 10 | 233678.1 | g2356198 | 2568 | 2897 |
| 10 | 233678.1 | 1684804T6 | 2577 | 2928 |
| 10 | 233678.1 | g2784196 | 2581 | 2968 |
| 10 | 233678.1 | 3236618H1 | 2583 | 2844 |
| 10 | 233678.1 | g2241760 | 2584 | 2968 |
| 10 | 233678.1 | g316793 | 2586 | 2979 |
| 10 | 233678.1 | g2816392 | 2588 | 2977 |
| 10 | 233678.1 | g1954535 | 2588 | 2782 |
| 10 | 233678.1 | g3700690 | 2586 | 2968 |
| 10 | 233678.1 | g1887793 | 2590 | 2968 |
| 10 | 233678.1 | 3643238H1 | 2593 | 2897 |
| 10 | 233678.1 | 3650438H1 | 2594 | 2894 |
| 10 | 233678.1 | 2466533H1 | 2612 | 2850 |
| 10 | 233678.1 | 2210370F6 | 1632 | 2079 |
| 10 | 233678.1 | 2210367H1 | 1632 | 1886 |
| 10 | 233678.1 | 4795285H1 | 1659 | 1948 |
| 10 | 233678.1 | g761628 | 1670 | 2092 |
| 10 | 233678.1 | 3803218H1 | 1669 | 1902 |
| 10 | 233678.1 | 4379275H1 | 1672 | 1949 |
| 10 | 233678.1 | 1606564H1 | 1682 | 1912 |
| 10 | 233678.1 | 5657623H1 | 1337 | 1607 |
| 10 | 233678.1 | 2312430H1 | 1358 | 1602 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 10 | 233678.1 | 3390808H1 | 1274 | 1553 |
| 10 | 233678.1 | 4171822H1 | 1390 | 1671 |
| 10 | 233678.1 | 4508866H1 | 1437 | 1711 |
| 10 | 233678.1 | 4505362H1 | 1437 | 1676 |
| 10 | 233678.1 | 632775H1 | 1451 | 1683 |
| 10 | 233678.1 | 2326782H1 | 1451 | 1677 |
| 10 | 233678.1 | 1684804H1 | 1281 | 1507 |
| 10 | 233678.1 | 1684804F6 | 1281 | 1743 |
| 10 | 233678.1 | 3948184H1 | 1457 | 1585 |
| 10 | 233678.1 | 3804480H1 | 1487 | 1706 |
| 10 | 233678.1 | 5546966H1 | 1486 | 1671 |
| 10 | 233678.1 | 5545829H1 | 1487 | 1700 |
| 10 | 233678.1 | 3088194H1 | 1498 | 1735 |
| 10 | 233678.1 | 3288867H1 | 1285 | 1546 |
| 10 | 233678.1 | 1250975F1 | 1577 | 2054 |
| 10 | 233678.1 | 1250975F6 | 1577 | 2051 |
| 10 | 233678.1 | 5812225H1 | 1317 | 1607 |
| 10 | 233678.1 | 5897556H1 | 1330 | 1555 |
| 10 | 233678.1 | 5893318H1 | 1330 | 1640 |
| 10 | 233678.1 | 5900195H1 | 1330 | 1548 |
| 10 | 233678.1 | 5897434H1 | 1330 | 1411 |
| 10 | 233678.1 | 5898715H1 | 1330 | 1634 |
| 10 | 233678.1 | 1250975H1 | 1577 | 1826 |
| 10 | 233678.1 | 5396479H1 | 1593 | 1852 |
| 10 | 233678.1 | 3797732H1 | 1604 | 1709 |
| 10 | 233678.1 | 5542957H1 | 1604 | 1815 |
| 10 | 233678.1 | 5558368H1 | 1609 | 1869 |
| 10 | 233678.1 | 1002256H1 | 1 | 246 |
| 10 | 233678.1 | 1002256R1 | 1 | 482 |
| 10 | 233678.1 | 1399306H1 | 74 | 326 |
| 10 | 233678.1 | 1399306F6 | 74 | 632 |
| 10 | 233678.1 | 1894055H1 | 182 | 422 |
| 10 | 233678.1 | 1270831F1 | 393 | 979 |
| 10 | 233678.1 | 1270831H1 | 393 | 667 |
| 10 | 233678.1 | 1003419R1 | 414 | 915 |
| 10 | 233678.1 | 1003419H1 | 414 | 608 |
| 10 | 233678.1 | 2518190F6 | 454 | 973 |
| 10 | 233678.1 | 2518190H1 | 454 | 706 |
| 10 | 233678.1 | 5659236H1 | 474 | 628 |
| 10 | 233678.1 | 3288389H1 | 1161 | 1413 |
| 10 | 233678.1 | 5005539H1 | 1189 | 1264 |
| 10 | 233678.1 | 5546782H1 | 1192 | 1398 |
| 10 | 233678.1 | 2809270H1 | 1239 | 1456 |
| 10 | 233678.1 | 2814466H1 | 762 | 1073 |
| 10 | 233678.1 | 5594064H1 | 496 | 757 |
| 10 | 233678.1 | g764169 | 598 | 923 |
| 10 | 233678.1 | 5546549H1 | 626 | 827 |
| 10 | 233678.1 | 3289542F6 | 652 | 1084 |
| 10 | 233678.1 | 3289542H1 | 652 | 893 |
| 10 | 233678.1 | 5544312H1 | 687 | 896 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 10 | 233678.1 | 2814466F6 | 762 | 1334 |
| 10 | 233678.1 | 2466560H1 | 766 | 1014 |
| 10 | 233678.1 | 2661506H1 | 769 | 1035 |
| 10 | 233678.1 | 2441878H1 | 884 | 988 |
| 10 | 233678.1 | 2437941H1 | 893 | 1114 |
| 10 | 233678.1 | 4785794H1 | 905 | 1166 |
| 10 | 233678.1 | 5574111H1 | 940 | 1157 |
| 10 | 233678.1 | 3964034H1 | 1006 | 1286 |
| 10 | 233678.1 | 4786766H1 | 1010 | 1266 |
| 10 | 233678.1 | 650941H1 | 1026 | 1285 |
| 10 | 233678.1 | 4601334H1 | 1061 | 1307 |
| 10 | 233678.1 | 3275906H1 | 1100 | 1352 |
| 10 | 233678.1 | 3274147F6 | 1101 | 1655 |
| 10 | 233678.1 | 3274147H1 | 1101 | 1348 |
| 10 | 233678.1 | g3163775 | 1145 | 1490 |
| 10 | 233678.1 | g4244100 | 2615 | 2969 |
| 10 | 233678.1 | g4329102 | 2618 | 2969 |
| 10 | 233678.1 | g4451360 | 2615 | 2971 |
| 10 | 233678.1 | g3802177 | 2627 | 2971 |
| 10 | 233678.1 | g3092982 | 2628 | 2969 |
| 10 | 233678.1 | g3921970 | 2626 | 2972 |
| 10 | 233678.1 | 3516016H1 | 2646 | 2918 |
| 10 | 233678.1 | 1377689H1 | 2654 | 2926 |
| 10 | 233678.1 | 4986111H1 | 2666 | 2956 |
| 10 | 233678.1 | g845966 | 2675 | 2968 |
| 10 | 233678.1 | g3777917 | 2683 | 2968 |
| 10 | 233678.1 | g761629 | 2702 | 2962 |
| 10 | 233678.1 | g3086793 | 2719 | 2968 |
| 10 | 233678.1 | 5273307H1 | 2727 | 2976 |
| 10 | 233678.1 | g1139236 | 2748 | 2972 |
| 10 | 233678.1 | 4987591H1 | 2793 | 2957 |
| 10 | 233678.1 | 4987593H1 | 2794 | 2968 |
| 10 | 233678.1 | 2424047H1 | 2804 | 2968 |
| 10 | 233678.1 | 6094676H1 | 2807 | 2968 |
| 10 | 233678.1 | 3083086H1 | 2842 | 2968 |
| 10 | 233678.1 | 3787162H1 | 2863 | 2968 |
| 10 | 233678.1 | 4901030H1 | 2902 | 2968 |
| 10 | 233678.1 | 5103514H1 | 2904 | 2974 |
| 10 | 233678.1 | 5060557H1 | 1720 | 2021 |
| 10 | 233678.1 | 5193174H1 | 1731 | 1877 |
| 10 | 233678.1 | 5003089H1 | 1750 | 2052 |
| 10 | 233678.1 | 3937103H1 | 1754 | 2052 |
| 10 | 233678.1 | 2210665H1 | 1754 | 1986 |
| 10 | 233678.1 | 4951636H2 | 1754 | 2033 |
| 10 | 233678.1 | 4589805H1 | 1754 | 1970 |
| 10 | 233678.1 | 2738891H1 | 1754 | 1967 |
| 10 | 233678.1 | 4793896H1 | 1754 | 2005 |
| 10 | 233678.1 | 1651922H1 | 1754 | 1961 |
| 10 | 233678.1 | 3121552H1 | 1754 | 2049 |
| 10 | 233678.1 | 1236003H1 | 1754 | 1975 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 10 | 233678.1 | 3361752H1 | 1754 | 2001 |
| 10 | 233678.1 | 3282866H1 | 1754 | 2003 |
| 10 | 233678.1 | 3114487H1 | 1754 | 2013 |
| 10 | 233678.1 | 1652026H1 | 1754 | 1943 |
| 10 | 233678.1 | 3804172H1 | 1809 | 2145 |
| 10 | 233678.1 | g2063923 | 1820 | 2231 |
| 10 | 233678.1 | 5191775H1 | 1834 | 1987 |
| 10 | 233678.1 | 1508970H1 | 1833 | 2054 |
| 10 | 233678.1 | 763920H1 | 1833 | 2139 |
| 10 | 233678.1 | 1803848H1 | 1839 | 1940 |
| 10 | 233678.1 | 2674708H1 | 1839 | 2112 |
| 11 | 312243.1 | 4605386H1 | 1343 | 1593 |
| 11 | 312243.1 | 4605386F7 | 1343 | 1784 |
| 11 | 312243.1 | 1483834H1 | 1462 | 1744 |
| 11 | 312243.1 | 4696494H1 | 1721 | 1986 |
| 11 | 312243.1 | 1823519H1 | 1734 | 1955 |
| 11 | 312243.1 | 1823519F6 | 1734 | 2013 |
| 11 | 312243.1 | 1696230H1 | 1749 | 1855 |
| 11 | 312243.1 | 1695671H1 | 1757 | 1986 |
| 11 | 312243.1 | 1696055H1 | 1757 | 1969 |
| 11 | 312243.1 | 1460083H1 | 1773 | 2007 |
| 11 | 312243.1 | 1823519T6 | 1787 | 2427 |
| 11 | 312243.1 | 3845858H1 | 1787 | 1988 |
| 11 | 312243.1 | 386870H1 | 1792 | 2067 |
| 11 | 312243.1 | g1515884 | 1886 | 2193 |
| 11 | 312243.1 | 2581154F6 | 21 | 197 |
| 11 | 312243.1 | 2581154H1 | 21 | 289 |
| 11 | 312243.1 | 694181H1 | 26 | 222 |
| 11 | 312243.1 | 3383640H1 | 33 | 277 |
| 11 | 312243.1 | 878997H1 | 250 | 387 |
| 11 | 312243.1 | 881475H1 | 252 | 495 |
| 11 | 312243.1 | 878997R1 | 252 | 816 |
| 11 | 312243.1 | 881475R6 | 252 | 705 |
| 11 | 312243.1 | 641653R6 | 480 | 1033 |
| 11 | 312243.1 | 641653H1 | 480 | 729 |
| 11 | 312243.1 | g2055007 | 674 | 772 |
| 11 | 312243.1 | g2063527 | 684 | 806 |
| 11 | 312243.1 | 1316282H1 | 912 | 1148 |
| 11 | 312243.1 | 2120964F6 | 1099 | 1478 |
| 11 | 312243.1 | 2120964H1 | 1099 | 1353 |
| 11 | 312243.1 | g4186493 | 1088 | 1442 |
| 11 | 312243.1 | g4451005 | 1162 | 1479 |
| 11 | 312243.1 | g2273834 | 1273 | 1458 |
| 11 | 312243.1 | 4605366H1 | 1343 | 1599 |
| 11 | 312243.1 | 4605386T7 | 1923 | 2422 |
| 11 | 312243.1 | 641653T6 | 1947 | 2441 |
| 11 | 312243.1 | 2120964T6 | 1957 | 2429 |
| 11 | 312243.1 | g4305745 | 2052 | 2366 |
| 11 | 312243.1 | 881475T6 | 2069 | 2428 |
| 11 | 312243.1 | g2138940 | 2087 | 2472 |

TABLE 4

| SEQ ID NO: | Template ID | Compon nt ID | Start | Stop |
|---|---|---|---|---|
| 11 | 312243.1 | 4601955H1 | 2114 | 2269 |
| 11 | 312243.1 | g3897513 | 2313 | 2467 |
| 11 | 312243.1 | g3281088 | 1 | 220 |
| 11 | 312243.1 | g4189325 | 1 | 177 |
| 11 | 312243.1 | g3419223 | 1 | 219 |
| 11 | 312243.1 | 3460792H1 | 1 | 206 |
| 12 | 425487.3 | g4187652 | 25 | 467 |
| 12 | 425487.3 | g3804148 | 48 | 413 |
| 12 | 425487.3 | g4265777 | 50 | 441 |
| 12 | 425487.3 | g1810525 | 25 | 243 |
| 12 | 425487.3 | g3075563 | 51 | 536 |
| 12 | 425487.3 | g1576978 | 582 | 740 |
| 12 | 425487.3 | g2276774 | 1 | 236 |
| 12 | 425487.3 | g1576933 | 1 | 386 |
| 12 | 425487.3 | g1860480 | 1 | 247 |
| 12 | 425487.3 | g3835346 | 1 | 498 |
| 12 | 425487.3 | g3839178 | 1 | 474 |
| 12 | 425487.3 | g3839255 | 1 | 504 |
| 12 | 425487.3 | g1920365 | 1 | 417 |
| 12 | 425487.3 | g4113340 | 25 | 317 |
| 12 | 425487.3 | 4845742H1 | 69 | 296 |
| 12 | 425487.3 | g2816806 | 66 | 402 |
| 12 | 425487.3 | g2881811 | 75 | 519 |
| 12 | 425487.3 | g2901201 | 84 | 461 |
| 12 | 425487.3 | g2819800 | 80 | 494 |
| 12 | 425487.3 | g2899484 | 91 | 540 |
| 12 | 425487.3 | g2240223 | 266 | 660 |
| 12 | 425487.3 | 5639438H1 | 531 | 769 |
| 18 | 222429.3 | 3135386H1 | 322 | 523 |
| 18 | 222429.3 | 4193895H1 | 745 | 1038 |
| 18 | 222429.3 | g3917627 | 750 | 967 |
| 18 | 222429.3 | 4773983H1 | 750 | 1030 |
| 18 | 222429.3 | 3110313H1 | 767 | 1070 |
| 18 | 222429.3 | 3818019H1 | 322 | 637 |
| 18 | 222429.3 | 5270075H1 | 770 | 1008 |
| 18 | 222429.3 | 2285210H1 | 324 | 583 |
| 18 | 222429.3 | 3451580H1 | 783 | 1039 |
| 18 | 222429.3 | 5690354H1 | 788 | 1045 |
| 18 | 222429.3 | 3368856H1 | 323 | 607 |
| 18 | 222429.3 | 5692143H1 | 796 | 1012 |
| 18 | 222429.3 | 1695061H1 | 796 | 886 |
| 18 | 222429.3 | 622658H1 | 803 | 1062 |
| 18 | 222429.3 | 4264839H1 | 814 | 968 |
| 18 | 222429.3 | 4382955H1 | 324 | 595 |
| 18 | 222429.3 | 2513281H1 | 325 | 584 |
| 18 | 222429.3 | g1395660 | 816 | 1293 |
| 18 | 222429.3 | 3182679H1 | 325 | 653 |
| 18 | 222429.3 | g2881072 | 817 | 1302 |
| 18 | 222429.3 | 1852753H1 | 328 | 622 |
| 18 | 222429.3 | 1852753F6 | 329 | 803 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 18 | 222429.3 | g4073351 | 823 | 1300 |
| 18 | 222429.3 | 3619548H1 | 332 | 599 |
| 18 | 222429.3 | 1772964H1 | 829 | 1110 |
| 18 | 222429.3 | 5016650H1 | 835 | 1123 |
| 18 | 222429.3 | 891033R1 | 835 | 967 |
| 18 | 222429.3 | 2782761H1 | 341 | 620 |
| 18 | 222429.3 | 4855161H1 | 347 | 487 |
| 18 | 222429.3 | 2802335H1 | 380 | 661 |
| 18 | 222429.3 | 4616684H1 | 380 | 653 |
| 18 | 222429.3 | g1395574 | 387 | 641 |
| 18 | 222429.3 | g2032347 | 393 | 677 |
| 18 | 222429.3 | 3553107H1 | 395 | 615 |
| 18 | 222429.3 | 4309262H1 | 423 | 747 |
| 18 | 222429.3 | 4721466T6 | 428 | 934 |
| 18 | 222429.3 | 3466526H1 | 435 | 706 |
| 18 | 222429.3 | 2316414H1 | 439 | 697 |
| 18 | 222429.3 | g3087119 | 484 | 970 |
| 18 | 222429.3 | 2859627H1 | 491 | 779 |
| 18 | 222429.3 | g3888504 | 540 | 971 |
| 18 | 222429.3 | 1266732H1 | 540 | 774 |
| 18 | 222429.3 | g3887405 | 551 | 869 |
| 18 | 222429.3 | g3801166 | 560 | 970 |
| 18 | 222429.3 | 759622R1 | 567 | 804 |
| 18 | 222429.3 | 759622H1 | 567 | 920 |
| 18 | 222429.3 | g1496653 | 586 | 967 |
| 18 | 222429.3 | 053812H1 | 589 | 802 |
| 18 | 222429.3 | 1822941H1 | 602 | 829 |
| 18 | 222429.3 | g3871758 | 609 | 967 |
| 18 | 222429.3 | g3888431 | 624 | 972 |
| 18 | 222429.3 | 5274484H1 | 626 | 878 |
| 18 | 222429.3 | g3802660 | 651 | 970 |
| 18 | 222429.3 | 1852753T6 | 655 | 1252 |
| 18 | 222429.3 | 6015049H1 | 672 | 969 |
| 18 | 222429.3 | 1430609H1 | 686 | 919 |
| 18 | 222429.3 | 798615H1 | 694 | 975 |
| 18 | 222429.3 | 3113611T6 | 698 | 1224 |
| 18 | 222429.3 | 1685142H1 | 742 | 976 |
| 18 | 222429.3 | 5900905H1 | 743 | 1032 |
| 18 | 222429.3 | 2078759H1 | 849 | 1127 |
| 18 | 222429.3 | 2043774H1 | 851 | 1124 |
| 18 | 222429.3 | g2985157 | 859 | 1298 |
| 18 | 222429.3 | 1402308H1 | 861 | 1118 |
| 18 | 222429.3 | g717242 | 882 | 971 |
| 18 | 222429.3 | g4109637 | 883 | 1293 |
| 18 | 222429.3 | g4110383 | 891 | 1293 |
| 18 | 222429.3 | g3178557 | 926 | 1301 |
| 18 | 222429.3 | 1463294H1 | 928 | 1142 |
| 18 | 222429.3 | 1463335H1 | 928 | 1120 |
| 18 | 222429.3 | 1463294T1 | 928 | 1248 |
| 18 | 222429.3 | g3871712 | 929 | 1292 |

TABLE 4

| SEQ ID NO: | Templat ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 18 | 222429.3 | g4307952 | 950 | 1295 |
| 18 | 222429.3 | 1940845T6 | 953 | 1251 |
| 18 | 222429.3 | 1940845R6 | 953 | 1294 |
| 18 | 222429.3 | 1940845H1 | 953 | 1181 |
| 18 | 222429.3 | 3730073H1 | 957 | 1278 |
| 18 | 222429.3 | g1218940 | 968 | 1293 |
| 18 | 222429.3 | g2575322 | 974 | 1293 |
| 18 | 222429.3 | 3726214H1 | 989 | 1294 |
| 18 | 222429.3 | 2872559H1 | 994 | 1288 |
| 18 | 222429.3 | 2875466H1 | 994 | 1134 |
| 18 | 222429.3 | 4247311H1 | 996 | 1144 |
| 18 | 222429.3 | 6007354H1 | 1016 | 1290 |
| 18 | 222429.3 | 2154383H1 | 1040 | 1293 |
| 18 | 222429.3 | 5680928H1 | 1043 | 1293 |
| 18 | 222429.3 | 214245H1 | 1081 | 1279 |
| 18 | 222429.3 | g1069690 | 1091 | 1208 |
| 18 | 222429.3 | g2212472 | 1098 | 1350 |
| 18 | 222429.3 | 839255H1 | 1103 | 1312 |
| 18 | 222429.3 | 839255R1 | 1103 | 1293 |
| 18 | 222429.3 | 3235786H1 | 1143 | 1252 |
| 18 | 222429.3 | 891762R1 | 1164 | 1300 |
| 18 | 222429.3 | 891762H1 | 1164 | 1300 |
| 18 | 222429.3 | 1849630H1 | 1184 | 1293 |
| 18 | 222429.3 | 1849630F6 | 1184 | 1284 |
| 18 | 222429.3 | 1849630T6 | 1192 | 1252 |
| 18 | 222429.3 | 4865293H1 | 1210 | 1293 |
| 18 | 222429.3 | 4721466F6 | 1 | 466 |
| 18 | 222429.3 | 4721466H1 | 1 | 271 |
| 18 | 222429.3 | 5638504H1 | 65 | 317 |
| 18 | 222429.3 | 5638852H1 | 65 | 316 |
| 18 | 222429.3 | 3076642H1 | 228 | 498 |
| 18 | 222429.3 | 3076642F6 | 228 | 487 |
| 18 | 222429.3 | 4940608H1 | 247 | 533 |
| 18 | 222429.3 | 5388153H1 | 267 | 412 |
| 18 | 222429.3 | 3750219H1 | 253 | 506 |
| 18 | 222429.3 | 3937336H1 | 253 | 527 |
| 18 | 222429.3 | 4122661H1 | 253 | 500 |
| 18 | 222429.3 | 3680329H1 | 253 | 572 |
| 18 | 222429.3 | 4943143H1 | 253 | 516 |
| 18 | 222429.3 | 4585314H1 | 254 | 510 |
| 18 | 222429.3 | 2732104H1 | 254 | 510 |
| 18 | 222429.3 | 2537505H1 | 257 | 492 |
| 18 | 222429.3 | g2015403 | 274 | 591 |
| 18 | 222429.3 | 4786776H1 | 274 | 526 |
| 18 | 222429.3 | 4045742H1 | 274 | 579 |
| 18 | 222429.3 | 3567962H1 | 274 | 529 |
| 18 | 222429.3 | 1386440H1 | 276 | 413 |
| 18 | 222429.3 | 5187636H1 | 275 | 560 |
| 18 | 222429.3 | 5197384H1 | 277 | 547 |
| 18 | 222429.3 | 1572627H1 | 277 | 491 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 18 | 222429.3 | 1572748H1 | 277 | 502 |
| 18 | 222429.3 | 2372113H1 | 279 | 527 |
| 18 | 222429.3 | 5187033H1 | 279 | 420 |
| 18 | 222429.3 | 2785651H1 | 279 | 563 |
| 18 | 222429.3 | 2727323H1 | 280 | 534 |
| 18 | 222429.3 | 3134034H1 | 280 | 554 |
| 18 | 222429.3 | 927455H1 | 280 | 558 |
| 18 | 222429.3 | 927455R1 | 280 | 894 |
| 18 | 222429.3 | 2596333H1 | 280 | 528 |
| 18 | 222429.3 | 2372113F6 | 280 | 797 |
| 18 | 222429.3 | 5847736H1 | 281 | 565 |
| 18 | 222429.3 | 2853635H1 | 281 | 538 |
| 18 | 222429.3 | 2854289H1 | 281 | 561 |
| 18 | 222429.3 | 2730059H1 | 281 | 540 |
| 18 | 222429.3 | 1268178F1 | 282 | 704 |
| 18 | 222429.3 | 2492274H1 | 282 | 517 |
| 18 | 222429.3 | 4124267H1 | 282 | 513 |
| 18 | 222429.3 | 2459370H1 | 282 | 511 |
| 18 | 222429.3 | 1268178H1 | 282 | 556 |
| 18 | 222429.3 | 6013641H1 | 282 | 518 |
| 18 | 222429.3 | 5378585H1 | 283 | 535 |
| 18 | 222429.3 | 3985882H1 | 274 | 384 |
| 18 | 222429.3 | 3870288H1 | 285 | 576 |
| 18 | 222429.3 | 4608143H1 | 283 | 534 |
| 18 | 222429.3 | 3983482H1 | 286 | 464 |
| 18 | 222429.3 | 5843644H1 | 286 | 511 |
| 18 | 222429.3 | 2605545H1 | 288 | 536 |
| 18 | 222429.3 | 3088525H1 | 288 | 577 |
| 18 | 222429.3 | 3218430H1 | 290 | 587 |
| 18 | 222429.3 | g1496652 | 291 | 538 |
| 18 | 222429.3 | 4248167H1 | 294 | 568 |
| 18 | 222429.3 | 4767534H1 | 294 | 424 |
| 18 | 222429.3 | 4174335H1 | 294 | 613 |
| 18 | 222429.3 | 3695889H1 | 295 | 582 |
| 18 | 222429.3 | 5592876H1 | 294 | 443 |
| 18 | 222429.3 | 2658509H1 | 296 | 546 |
| 18 | 222429.3 | 730756H1 | 296 | 565 |
| 18 | 222429.3 | 4613408H1 | 297 | 505 |
| 18 | 222429.3 | 264145H1 | 297 | 616 |
| 18 | 222429.3 | 5810372H1 | 299 | 665 |
| 18 | 222429.3 | 3319788H1 | 300 | 476 |
| 18 | 222429.3 | 4202422H1 | 301 | 596 |
| 18 | 222429.3 | 4613487H1 | 306 | 507 |
| 18 | 222429.3 | g712550 | 306 | 583 |
| 18 | 222429.3 | 5090055H1 | 307 | 567 |
| 18 | 222429.3 | 2823320H1 | 309 | 631 |
| 18 | 222429.3 | 2529472H1 | 311 | 578 |
| 18 | 222429.3 | 4850016H1 | 312 | 590 |
| 18 | 222429.3 | 4606729H1 | 313 | 569 |
| 18 | 222429.3 | 4608174H1 | 313 | 567 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 18 | 222429.3 | 4334782H1 | 314 | 584 |
| 18 | 222429.3 | 3508043H1 | 314 | 609 |
| 18 | 222429.3 | 4638791H1 | 316 | 578 |
| 18 | 222429.3 | 2941851H1 | 316 | 594 |
| 18 | 222429.3 | 2533245H1 | 317 | 556 |
| 18 | 222429.3 | 033345H1 | 317 | 444 |
| 18 | 222429.3 | 2888083H1 | 317 | 573 |
| 18 | 222429.3 | 2766782H1 | 318 | 591 |
| 18 | 222429.3 | g1069689 | 321 | 669 |
| 18 | 222429.3 | 2514805H1 | 321 | 656 |
| 19 | 366739.2 | 3793584H1 | 1 | 276 |
| 19 | 366739.2 | 3093259F6 | 1 | 396 |
| 19 | 366739.2 | 3093259H1 | 1 | 272 |
| 19 | 366739.2 | 5586972H1 | 28 | 288 |
| 19 | 366739.2 | 1562228H1 | 28 | 244 |
| 19 | 366739.2 | g1976820 | 31 | 385 |
| 19 | 366739.2 | 5472436H1 | 30 | 254 |
| 19 | 366739.2 | g1974601 | 38 | 290 |
| 19 | 366739.2 | 3536311H1 | 63 | 278 |
| 19 | 366739.2 | g1301142 | 203 | 492 |
| 19 | 366739.2 | 2996019H1 | 226 | 495 |
| 19 | 366739.2 | 1291844F6 | 265 | 754 |
| 19 | 366739.2 | 1291844F1 | 265 | 757 |
| 19 | 366739.2 | 1291844H1 | 265 | 495 |
| 19 | 366739.2 | 5107426H1 | 323 | 395 |
| 19 | 366739.2 | 6077830H1 | 331 | 653 |
| 19 | 366739.2 | 2690675H1 | 335 | 549 |
| 19 | 366739.2 | 3673120H1 | 341 | 654 |
| 19 | 366739.2 | 4626753H1 | 343 | 626 |
| 19 | 366739.2 | 2814563H1 | 344 | 559 |
| 19 | 366739.2 | 5019593H1 | 348 | 607 |
| 19 | 366739.2 | 5591354H1 | 350 | 602 |
| 19 | 366739.2 | 4337887H1 | 348 | 533 |
| 19 | 366739.2 | 2830045H1 | 350 | 618 |
| 19 | 366739.2 | 3614922H1 | 351 | 637 |
| 19 | 366739.2 | 3034106H1 | 352 | 640 |
| 19 | 366739.2 | 4765938H1 | 352 | 635 |
| 19 | 366739.2 | 3375151H1 | 354 | 621 |
| 19 | 366739.2 | 804576H1 | 357 | 612 |
| 19 | 366739.2 | 2188984H1 | 360 | 622 |
| 19 | 366739.2 | 157111H1 | 361 | 569 |
| 19 | 366739.2 | 5391316H1 | 370 | 647 |
| 19 | 366739.2 | 808513H1 | 372 | 654 |
| 19 | 366739.2 | 1966930H1 | 381 | 617 |
| 19 | 366739.2 | 1966930R6 | 381 | 755 |
| 19 | 366739.2 | 5194739H1 | 431 | 658 |
| 19 | 366739.2 | 4946548H1 | 431 | 604 |
| 19 | 366739.2 | 4768533F6 | 455 | 865 |
| 19 | 366739.2 | 4768533H1 | 455 | 739 |
| 19 | 366739.2 | g778805 | 457 | 743 |

TABLE 4

| SEQ ID NO: | Template ID | Compon nt ID | Start | Stop |
|---|---|---|---|---|
| 19 | 366739.2 | g1921131 | 463 | 977 |
| 19 | 366739.2 | 4353206H2 | 492 | 757 |
| 19 | 366739.2 | 5325529H1 | 501 | 794 |
| 19 | 366739.2 | 5322285H1 | 501 | 774 |
| 19 | 366739.2 | 5322857H1 | 502 | 752 |
| 19 | 366739.2 | 869959H1 | 564 | 808 |
| 19 | 366739.2 | 875716R6 | 564 | 1081 |
| 19 | 366739.2 | 875716H1 | 564 | 855 |
| 19 | 366739.2 | 875716R1 | 564 | 1203 |
| 19 | 366739.2 | 3221359H1 | 601 | 939 |
| 19 | 366739.2 | 3869036H1 | 630 | 910 |
| 19 | 366739.2 | 3353207H1 | 645 | 955 |
| 19 | 366739.2 | g2016781 | 644 | 916 |
| 19 | 366739.2 | 4587185H1 | 645 | 907 |
| 19 | 366739.2 | 5406614H1 | 653 | 850 |
| 19 | 366739.2 | g1300754 | 650 | 1123 |
| 19 | 366739.2 | 370417H1 | 713 | 1027 |
| 19 | 366739.2 | 1265285H1 | 720 | 861 |
| 19 | 366739.2 | 1265285R1 | 720 | 1186 |
| 19 | 366739.2 | 1265102H1 | 720 | 891 |
| 19 | 366739.2 | 3946741H1 | 742 | 1024 |
| 19 | 366739.2 | 3603130H1 | 771 | 1093 |
| 19 | 366739.2 | 5472209H1 | 846 | 1043 |
| 19 | 366739.2 | 3717241H1 | 870 | 1172 |
| 19 | 366739.2 | 1291844T6 | 871 | 1416 |
| 19 | 366739.2 | 2608544H1 | 898 | 1103 |
| 19 | 366739.2 | g922177 | 903 | 1148 |
| 19 | 366739.2 | 251965H1 | 925 | 1273 |
| 19 | 366739.2 | 3470188H1 | 935 | 1210 |
| 19 | 366739.2 | g2838122 | 936 | 1451 |
| 19 | 366739.2 | 1636138F6 | 963 | 1441 |
| 19 | 366739.2 | 1636138H1 | 963 | 1184 |
| 19 | 366739.2 | 923492H1 | 968 | 1270 |
| 19 | 366739.2 | 3811371H1 | 977 | 1294 |
| 19 | 366739.2 | g3430481 | 978 | 1454 |
| 19 | 366739.2 | g4282887 | 982 | 1458 |
| 19 | 366739.2 | g4372283 | 986 | 1454 |
| 19 | 366739.2 | 5138130H1 | 1015 | 1307 |
| 19 | 366739.2 | g2958044 | 1019 | 1453 |
| 19 | 366739.2 | 1636138T6 | 1035 | 1416 |
| 19 | 366739.2 | g2397938 | 1039 | 1452 |
| 19 | 366739.2 | 1415729H1 | 1048 | 1289 |
| 19 | 366739.2 | 676581H1 | 1055 | 1273 |
| 19 | 366739.2 | 3093259T6 | 1056 | 1412 |
| 19 | 366739.2 | g1921132 | 1059 | 1465 |
| 19 | 366739.2 | 337785H1 | 1075 | 1288 |
| 19 | 366739.2 | 3773169H1 | 1088 | 1367 |
| 19 | 366739.2 | g2820846 | 1095 | 1453 |
| 19 | 366739.2 | g2322961 | 1109 | 1457 |
| 19 | 366739.2 | g3923575 | 1124 | 1453 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 19 | 366739.2 | 5399095H1 | 1135 | 1273 |
| 19 | 366739.2 | g1266798 | 1140 | 1459 |
| 19 | 366739.2 | g2883870 | 1143 | 1434 |
| 19 | 366739.2 | g4325983 | 1152 | 1458 |
| 19 | 366739.2 | g4325893 | 1154 | 1458 |
| 19 | 366739.2 | 1223335T1 | 1154 | 1414 |
| 19 | 366739.2 | 1223335H1 | 1154 | 1434 |
| 19 | 366739.2 | g3959938 | 1177 | 1453 |
| 19 | 366739.2 | g1970868 | 1179 | 1468 |
| 19 | 366739.2 | 2228180H1 | 1185 | 1460 |
| 19 | 366739.2 | g2574692 | 1223 | 1462 |
| 19 | 366739.2 | 2505795H1 | 1237 | 1456 |
| 19 | 366739.2 | g2877025 | 1242 | 1453 |
| 19 | 366739.2 | g4268440 | 1259 | 1453 |
| 19 | 366739.2 | g2575341 | 1264 | 1453 |
| 19 | 366739.2 | g778806 | 1268 | 1452 |
| 19 | 366739.2 | 1772854R6 | 1273 | 1453 |
| 19 | 366739.2 | 1772849H1 | 1273 | 1453 |
| 19 | 366739.2 | g1264282 | 1343 | 1473 |
| 20 | 474635.6 | 340786H1 | 2181 | 2422 |
| 20 | 474635.6 | 2242239H1 | 2179 | 2347 |
| 20 | 474635.6 | g991123 | 2179 | 2514 |
| 20 | 474635.6 | 2234442H1 | 2179 | 2397 |
| 20 | 474635.6 | g1056895 | 2180 | 2498 |
| 20 | 474635.6 | g959046 | 2182 | 2471 |
| 20 | 474635.6 | g1004715 | 2183 | 2527 |
| 20 | 474635.6 | 4996329T6 | 2189 | 2625 |
| 20 | 474635.6 | 4746496H1 | 2192 | 2451 |
| 20 | 474635.6 | g3737429 | 2197 | 2728 |
| 20 | 474635.6 | g1039999 | 2210 | 2512 |
| 20 | 474635.6 | g1081479 | 2210 | 2391 |
| 20 | 474635.6 | 2204795T6 | 2217 | 2671 |
| 20 | 474635.6 | g3736649 | 2220 | 2704 |
| 20 | 474635.6 | 2242239T6 | 2220 | 2816 |
| 20 | 474635.6 | 3733918T6 | 2230 | 2786 |
| 20 | 474635.6 | g3742588 | 2229 | 2728 |
| 20 | 474635.6 | g1014336 | 2235 | 2566 |
| 20 | 474635.6 | 3638137H1 | 2230 | 2511 |
| 20 | 474635.6 | g835640 | 2232 | 2567 |
| 20 | 474635.6 | 1666516H1 | 2236 | 2449 |
| 20 | 474635.6 | 2083270H1 | 2262 | 2567 |
| 20 | 474635.6 | 1761002H1 | 2284 | 2581 |
| 20 | 474635.6 | 612373H1 | 2291 | 2581 |
| 20 | 474635.6 | 665014H1 | 2291 | 2558 |
| 20 | 474635.6 | g867547 | 2291 | 2662 |
| 20 | 474635.6 | g1986029 | 2312 | 2772 |
| 20 | 474635.6 | 3000193H1 | 2318 | 2627 |
| 20 | 474635.6 | 461079H1 | 2328 | 2588 |
| 20 | 474635.6 | g1274411 | 2332 | 2863 |
| 20 | 474635.6 | 2237019H1 | 2332 | 2557 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 20 | 474635.6 | g856812 | 2334 | 2693 |
| 20 | 474635.6 | 5275784H1 | 2343 | 2627 |
| 20 | 474635.6 | 4085512H1 | 2393 | 2714 |
| 20 | 474635.6 | g1240371 | 2407 | 2873 |
| 20 | 474635.6 | g848488 | 2408 | 2767 |
| 20 | 474635.6 | g897948 | 2408 | 2722 |
| 20 | 474635.6 | g848512 | 2408 | 2850 |
| 20 | 474635.6 | g1444420 | 2410 | 2880 |
| 20 | 474635.6 | 665169H1 | 2419 | 2677 |
| 20 | 474635.6 | 2090332H1 | 2439 | 2728 |
| 20 | 474635.6 | g1004911 | 2454 | 2877 |
| 20 | 474635.6 | g2900276 | 2464 | 2950 |
| 20 | 474635.6 | g3848664 | 2464 | 2910 |
| 20 | 474635.6 | 5432669H1 | 2472 | 2741 |
| 20 | 474635.6 | 3134278H1 | 2486 | 2792 |
| 20 | 474635.6 | g4174578 | 2487 | 2999 |
| 20 | 474635.6 | g3433111 | 2489 | 2991 |
| 20 | 474635.6 | g3419207 | 2491 | 3017 |
| 20 | 474635.6 | 4087752H1 | 2502 | 2814 |
| 20 | 474635.6 | 5327883H1 | 2520 | 2806 |
| 20 | 474635.6 | 5326114H1 | 2523 | 2818 |
| 20 | 474635.6 | 5326214H1 | 2524 | 2806 |
| 20 | 474635.6 | g3229542 | 2527 | 3017 |
| 20 | 474635.6 | 2757338F6 | 2528 | 3008 |
| 20 | 474635.6 | g3405868 | 938 | 1282 |
| 20 | 474635.6 | 572947H1 | 944 | 1190 |
| 20 | 474635.6 | 3002964H1 | 961 | 1013 |
| 20 | 474635.6 | g982216 | 976 | 1338 |
| 20 | 474635.6 | 2309094H1 | 1029 | 1217 |
| 20 | 474635.6 | 4638158H1 | 1036 | 1287 |
| 20 | 474635.6 | 1912857H1 | 1042 | 1285 |
| 20 | 474635.6 | g916500 | 1047 | 1247 |
| 20 | 474635.6 | 3633880H1 | 1055 | 1354 |
| 20 | 474635.6 | 3634680H1 | 1055 | 1325 |
| 20 | 474635.6 | 4546165H1 | 1074 | 1351 |
| 20 | 474635.6 | g1276377 | 1073 | 1325 |
| 20 | 474635.6 | 5677925H1 | 1154 | 1382 |
| 20 | 474635.6 | 4663039H1 | 1163 | 1425 |
| 20 | 474635.6 | 3511926H1 | 1174 | 1426 |
| 20 | 474635.6 | 4996329H1 | 1324 | 1605 |
| 20 | 474635.6 | 4996329F6 | 1324 | 1683 |
| 20 | 474635.6 | 4740909H1 | 1330 | 1603 |
| 20 | 474635.6 | 647910H1 | 1344 | 1623 |
| 20 | 474635.6 | 1715955F6 | 1431 | 1896 |
| 20 | 474635.6 | 3281456H1 | 1463 | 1723 |
| 20 | 474635.6 | 2204795F6 | 1494 | 2015 |
| 20 | 474635.6 | 2204795H1 | 1494 | 1739 |
| 20 | 474635.6 | 4546440H1 | 1504 | 1781 |
| 20 | 474635.6 | 1669175H1 | 1524 | 1751 |
| 20 | 474635.6 | 1443714F6 | 1644 | 2033 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 20 | 474635.6 | 1443714H1 | 1644 | 1900 |
| 20 | 474635.6 | 1715955H1 | 1680 | 1896 |
| 20 | 474635.6 | 4088721H1 | 1699 | 1960 |
| 20 | 474635.6 | 4049359H1 | 1718 | 1989 |
| 20 | 474635.6 | 3322874H1 | 1720 | 1987 |
| 20 | 474635.6 | 3733918F6 | 1806 | 2215 |
| 20 | 474635.6 | 4411044H1 | 1832 | 2038 |
| 20 | 474635.6 | 5173224H1 | 1866 | 2144 |
| 20 | 474635.6 | 5098372H1 | 1930 | 2150 |
| 20 | 474635.6 | 4412048H1 | 1933 | 2137 |
| 20 | 474635.6 | 1265323R1 | 1935 | 2488 |
| 20 | 474635.6 | 1265323H1 | 1935 | 2151 |
| 20 | 474635.6 | 376565H1 | 1961 | 2151 |
| 20 | 474635.6 | 1630741H1 | 1961 | 2147 |
| 20 | 474635.6 | 1631939H1 | 1961 | 2151 |
| 20 | 474635.6 | 6102550H1 | 1966 | 2151 |
| 20 | 474635.6 | 6102850H1 | 1966 | 2154 |
| 20 | 474635.6 | g958816 | 2172 | 2378 |
| 20 | 474635.6 | g1047675 | 2173 | 2505 |
| 20 | 474635.6 | g965279 | 2173 | 2563 |
| 20 | 474635.6 | g3098969 | 2173 | 2510 |
| 20 | 474635.6 | g1202194 | 2174 | 2528 |
| 20 | 474635.6 | g1099631 | 2174 | 2567 |
| 20 | 474635.6 | g1047662 | 2174 | 2514 |
| 20 | 474635.6 | g1228299 | 2176 | 2306 |
| 20 | 474635.6 | g1193583 | 2176 | 2464 |
| 20 | 474635.6 | g2912515 | 2176 | 2667 |
| 20 | 474635.6 | 2242239F6 | 2176 | 2646 |
| 20 | 474635.6 | g1211570 | 2176 | 2487 |
| 20 | 474635.6 | g1225734 | 2176 | 2491 |
| 20 | 474635.6 | g1226785 | 2176 | 2462 |
| 20 | 474635.6 | g1043793 | 2176 | 2392 |
| 20 | 474635.6 | g1162551 | 2176 | 2505 |
| 20 | 474635.6 | 1725002H1 | 1 | 207 |
| 20 | 474635.6 | 1726247H1 | 1 | 216 |
| 20 | 474635.6 | 3495066H1 | 1 | 301 |
| 20 | 474635.6 | 1727369F6 | 1 | 387 |
| 20 | 474635.6 | 1727369H1 | 1 | 249 |
| 20 | 474635.6 | 3109953H1 | 10 | 305 |
| 20 | 474635.6 | 2362517H1 | 16 | 279 |
| 20 | 474635.6 | 2361835H1 | 16 | 270 |
| 20 | 474635.6 | 3617437H1 | 58 | 382 |
| 20 | 474635.6 | 2659652H1 | 91 | 348 |
| 20 | 474635.6 | 3734189H1 | 92 | 393 |
| 20 | 474635.6 | 2679260H1 | 96 | 409 |
| 20 | 474635.6 | 1466462H1 | 96 | 295 |
| 20 | 474635.6 | 3563056H1 | 96 | 415 |
| 20 | 474635.6 | 2235355H1 | 100 | 364 |
| 20 | 474635.6 | 2202076H1 | 99 | 367 |
| 20 | 474635.6 | 3129502H1 | 101 | 436 |

TABLE 4

| SEQ ID NO: | T mplate ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 20 | 474635.6 | 2051824H1 | 118 | 293 |
| 20 | 474635.6 | 3633032H1 | 122 | 294 |
| 20 | 474635.6 | 3150266H1 | 126 | 379 |
| 20 | 474635.6 | 2051825H1 | 132 | 414 |
| 20 | 474635.6 | g1227685 | 128 | 482 |
| 20 | 474635.6 | 4288817H1 | 146 | 437 |
| 20 | 474635.6 | 4860794H1 | 155 | 246 |
| 20 | 474635.6 | 174879H1 | 161 | 364 |
| 20 | 474635.6 | 5700836H1 | 163 | 471 |
| 20 | 474635.6 | 5539631H2 | 182 | 413 |
| 20 | 474635.6 | 5206045H1 | 220 | 491 |
| 20 | 474635.6 | 375911R6 | 244 | 566 |
| 20 | 474635.6 | 375911H1 | 244 | 536 |
| 20 | 474635.6 | g657023 | 255 | 500 |
| 20 | 474635.6 | 3939757H1 | 275 | 433 |
| 20 | 474635.6 | 3939749H1 | 276 | 433 |
| 20 | 474635.6 | g704964 | 290 | 589 |
| 20 | 474635.6 | g1013184 | 290 | 695 |
| 20 | 474635.6 | g1472478 | 326 | 795 |
| 20 | 474635.6 | 1220725H1 | 355 | 600 |
| 20 | 474635.6 | 4980839H1 | 358 | 659 |
| 20 | 474635.6 | 375911T6 | 370 | 902 |
| 20 | 474635.6 | 3675542H1 | 371 | 682 |
| 20 | 474635.6 | 3669542H1 | 371 | 695 |
| 20 | 474635.6 | 3671542H1 | 371 | 558 |
| 20 | 474635.6 | g1447913 | 383 | 835 |
| 20 | 474635.6 | 1282181H1 | 391 | 534 |
| 20 | 474635.6 | g2020599 | 418 | 783 |
| 20 | 474635.6 | 4133026H2 | 447 | 725 |
| 20 | 474635.6 | 2236848H1 | 450 | 698 |
| 20 | 474635.6 | g1278197 | 473 | 1017 |
| 20 | 474635.6 | g1941907 | 514 | 972 |
| 20 | 474635.6 | g2695124 | 514 | 1013 |
| 20 | 474635.6 | 1470350F6 | 554 | 998 |
| 20 | 474635.6 | 578500H1 | 557 | 758 |
| 20 | 474635.6 | g4371917 | 558 | 1013 |
| 20 | 474635.6 | g1447816 | 558 | 1013 |
| 20 | 474635.6 | g2702599 | 559 | 1013 |
| 20 | 474635.6 | 505476H1 | 560 | 793 |
| 20 | 474635.6 | 2086083H1 | 579 | 884 |
| 20 | 474635.6 | g3430549 | 602 | 1020 |
| 20 | 474635.6 | g1671099 | 609 | 1029 |
| 20 | 474635.6 | g1472421 | 658 | 1013 |
| 20 | 474635.6 | g656857 | 700 | 1013 |
| 20 | 474635.6 | 3381660H1 | 715 | 902 |
| 20 | 474635.6 | 4270822H1 | 742 | 1016 |
| 20 | 474635.6 | g1977606 | 753 | 1013 |
| 20 | 474635.6 | 4594512H1 | 755 | 1013 |
| 20 | 474635.6 | g944107 | 761 | 1016 |
| 20 | 474635.6 | 1470350H1 | 802 | 998 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 20 | 474635.6 | 2362517T6 | 802 | 970 |
| 20 | 474635.6 | g1011912 | 812 | 1004 |
| 20 | 474635.6 | g1201979 | 871 | 1013 |
| 20 | 474635.6 | 4412657H1 | 883 | 969 |
| 20 | 474635.6 | g916501 | 928 | 1234 |
| 20 | 474635.6 | g1238628 | 931 | 1126 |
| 20 | 474635.6 | g2057226 | 934 | 1034 |
| 20 | 474635.6 | 3450474H1 | 938 | 1128 |
| 20 | 474635.6 | g2022938 | 938 | 1147 |
| 20 | 474635.6 | g982171 | 939 | 1266 |
| 20 | 474635.6 | 2757338H1 | 2528 | 2825 |
| 20 | 474635.6 | 2757338R6 | 2528 | 2995 |
| 20 | 474635.6 | 5425690H1 | 2531 | 2796 |
| 20 | 474635.6 | g1046524 | 2535 | 2900 |
| 20 | 474635.6 | g3134715 | 2534 | 2985 |
| 20 | 474635.6 | 380176H1 | 2537 | 2654 |
| 20 | 474635.6 | g1186355 | 2560 | 2772 |
| 20 | 474635.6 | g856701 | 2582 | 2772 |
| 20 | 474635.6 | 1795930H1 | 2583 | 2701 |
| 20 | 474635.6 | g651862 | 2589 | 2849 |
| 20 | 474635.6 | g651879 | 2589 | 2896 |
| 20 | 474635.6 | g3076096 | 2594 | 2992 |
| 20 | 474635.6 | g4078849 | 2605 | 2999 |
| 20 | 474635.6 | g835594 | 2606 | 2985 |
| 20 | 474635.6 | g1081771 | 2618 | 2893 |
| 20 | 474635.6 | g1266372 | 2625 | 2900 |
| 20 | 474635.6 | 971739H1 | 2631 | 2952 |
| 20 | 474635.6 | 1808492H1 | 2639 | 2869 |
| 20 | 474635.6 | 649261H1 | 2645 | 2940 |
| 20 | 474635.6 | g867525 | 2652 | 2949 |
| 20 | 474635.6 | 1352072H1 | 2682 | 2960 |
| 20 | 474635.6 | g848398 | 2688 | 2993 |
| 20 | 474635.6 | g3754493 | 2700 | 2999 |
| 20 | 474635.6 | g2839098 | 2719 | 3187 |
| 20 | 474635.6 | g848419 | 2728 | 3010 |
| 20 | 474635.6 | g1524564 | 2740 | 3015 |
| 20 | 474635.6 | g806064 | 2765 | 2972 |
| 20 | 474635.6 | 1402667H1 | 2791 | 3002 |
| 20 | 474635.6 | g2752910 | 2841 | 2938 |
| 20 | 474635.6 | g715956 | 2849 | 2968 |
| 20 | 474635.6 | g3428030 | 2948 | 2997 |
| 24 | 411449.2 | 3292975H1 | 654 | 917 |
| 24 | 411449.2 | g1277242 | 663 | 1052 |
| 24 | 411449.2 | 1355750H1 | 662 | 944 |
| 24 | 411449.2 | 1355750F6 | 662 | 1150 |
| 24 | 411449.2 | g1443735 | 707 | 1043 |
| 24 | 411449.2 | 1984010H1 | 722 | 990 |
| 24 | 411449.2 | g1243124 | 747 | 1012 |
| 24 | 411449.2 | g1243130 | 747 | 939 |
| 24 | 411449.2 | g1243104 | 748 | 938 |

TABLE 4

| SEQ ID NO: | Templat ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 24 | 411449.2 | g1243131 | 773 | 923 |
| 24 | 411449.2 | g841841 | 823 | 1176 |
| 24 | 411449.2 | g1635822 | 830 | 1026 |
| 24 | 411449.2 | 677686H1 | 890 | 1140 |
| 24 | 411449.2 | 2896619H1 | 891 | 1077 |
| 24 | 411449.2 | g3933055 | 644 | 1016 |
| 24 | 411449.2 | 1689376H1 | 898 | 1105 |
| 24 | 411449.2 | 3783111H1 | 908 | 1221 |
| 24 | 411449.2 | 4796764H1 | 515 | 795 |
| 24 | 411449.2 | 4796772H1 | 515 | 795 |
| 24 | 411449.2 | g2037311 | 533 | 814 |
| 24 | 411449.2 | g2819773 | 551 | 815 |
| 24 | 411449.2 | 057539H1 | 507 | 708 |
| 24 | 411449.2 | g3098856 | 560 | 922 |
| 24 | 411449.2 | g1125331 | 560 | 950 |
| 24 | 411449.2 | g1386245 | 583 | 970 |
| 24 | 411449.2 | 3323088H1 | 586 | 854 |
| 24 | 411449.2 | g2816548 | 607 | 1022 |
| 24 | 411449.2 | 3660962H1 | 612 | 867 |
| 24 | 411449.2 | g1887649 | 615 | 946 |
| 24 | 411449.2 | g1489541 | 618 | 866 |
| 24 | 411449.2 | g1489523 | 619 | 965 |
| 24 | 411449.2 | g4327095 | 909 | 1337 |
| 24 | 411449.2 | 5275975H1 | 911 | 1077 |
| 24 | 411449.2 | 2682468H1 | 941 | 1157 |
| 24 | 411449.2 | 2682429H1 | 943 | 1218 |
| 24 | 411449.2 | 3448516T6 | 966 | 1496 |
| 24 | 411449.2 | 5517642H1 | 966 | 1232 |
| 24 | 411449.2 | 3723582H1 | 970 | 1267 |
| 24 | 411449.2 | 589712H1 | 978 | 1222 |
| 24 | 411449.2 | 589712R1 | 978 | 1519 |
| 24 | 411449.2 | 2613128H1 | 1015 | 1252 |
| 24 | 411449.2 | 1355750T6 | 1052 | 1482 |
| 24 | 411449.2 | 3820979H1 | 1057 | 1340 |
| 24 | 411449.2 | 1811236H1 | 1068 | 1326 |
| 24 | 411449.2 | 1811236F6 | 1068 | 1562 |
| 24 | 411449.2 | 1811236T6 | 1083 | 1726 |
| 24 | 411449.2 | g4033830 | 1106 | 1519 |
| 24 | 411449.2 | g3917075 | 1106 | 1519 |
| 24 | 411449.2 | g3917072 | 1108 | 1519 |
| 24 | 411449.2 | 3297093H1 | 1117 | 1204 |
| 24 | 411449.2 | 4513565H1 | 1174 | 1440 |
| 24 | 411449.2 | 4798512H1 | 1190 | 1460 |
| 24 | 411449.2 | 2645110H1 | 1195 | 1450 |
| 24 | 411449.2 | 2096680R6 | 1216 | 1644 |
| 24 | 411449.2 | 2096680H1 | 1216 | 1462 |
| 24 | 411449.2 | 321613H1 | 1233 | 1492 |
| 24 | 411449.2 | 2995437H1 | 1251 | 1538 |
| 24 | 411449.2 | g4371719 | 1378 | 1775 |
| 24 | 411449.2 | 2096680T6 | 1380 | 1715 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 24 | 411449.2 | 4653462H1 | 1411 | 1685 |
| 24 | 411449.2 | g1489542 | 1416 | 1765 |
| 24 | 411449.2 | 4144044H1 | 1422 | 1709 |
| 24 | 411449.2 | g1224160 | 1425 | 1770 |
| 24 | 411449.2 | g1224144 | 1433 | 1770 |
| 24 | 411449.2 | g1489524 | 1473 | 1766 |
| 24 | 411449.2 | g769281 | 1485 | 1823 |
| 24 | 411449.2 | g2955239 | 1501 | 1777 |
| 24 | 411449.2 | g4078417 | 1503 | 1770 |
| 24 | 411449.2 | g1224166 | 1517 | 1770 |
| 24 | 411449.2 | g1224143 | 1529 | 1770 |
| 24 | 411449.2 | 1972386H1 | 1531 | 1804 |
| 24 | 411449.2 | g4084798 | 1534 | 1954 |
| 24 | 411449.2 | g1224165 | 1560 | 1770 |
| 24 | 411449.2 | g784201 | 1564 | 1798 |
| 24 | 411449.2 | g2657184 | 1564 | 1868 |
| 24 | 411449.2 | 5919979H1 | 1256 | 1559 |
| 24 | 411449.2 | 008248H1 | 1258 | 1569 |
| 24 | 411449.2 | g889333 | 1271 | 1666 |
| 24 | 411449.2 | g1139751 | 1287 | 1746 |
| 24 | 411449.2 | 589712F1 | 1305 | 1770 |
| 24 | 411449.2 | 3659604H1 | 1316 | 1588 |
| 24 | 411449.2 | 2598057H1 | 1327 | 1458 |
| 24 | 411449.2 | g3178153 | 1345 | 1772 |
| 24 | 411449.2 | g4187548 | 1373 | 1774 |
| 24 | 411449.2 | g574838 | 1617 | 1824 |
| 24 | 411449.2 | 3778628H1 | 1645 | 1948 |
| 24 | 411449.2 | g3785651 | 1656 | 1770 |
| 24 | 411449.2 | 763788H1 | 1696 | 1765 |
| 24 | 411449.2 | 1490003H1 | 1788 | 2054 |
| 24 | 411449.2 | g2000780 | 1844 | 2185 |
| 24 | 411449.2 | 4704775H1 | 1862 | 2105 |
| 24 | 411449.2 | g3921575 | 1873 | 2274 |
| 24 | 411449.2 | g3231159 | 1901 | 2276 |
| 24 | 411449.2 | g2222962 | 1902 | 2260 |
| 24 | 411449.2 | g2222973 | 1902 | 2258 |
| 24 | 411449.2 | 638241H1 | 1908 | 2160 |
| 24 | 411449.2 | 001302H1 | 1908 | 2236 |
| 24 | 411449.2 | 1625423F6 | 1911 | 2197 |
| 24 | 411449.2 | 1625423H1 | 1911 | 2113 |
| 24 | 411449.2 | 491356H1 | 1923 | 2185 |
| 24 | 411449.2 | g1964447 | 1927 | 2193 |
| 24 | 411449.2 | g1963729 | 1927 | 2263 |
| 24 | 411449.2 | g1379590 | 1936 | 2272 |
| 24 | 411449.2 | g1472323 | 1948 | 2197 |
| 24 | 411449.2 | 2399751H1 | 1951 | 2192 |
| 24 | 411449.2 | 4856606H1 | 1973 | 2197 |
| 24 | 411449.2 | 995171H1 | 1978 | 2209 |
| 24 | 411449.2 | g1472316 | 1979 | 2197 |
| 24 | 411449.2 | 5194364H2 | 1982 | 2194 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 24 | 411449.2 | 3451002H1 | 2043 | 2168 |
| 24 | 411449.2 | g4524789 | 2083 | 2553 |
| 24 | 411449.2 | g2159625 | 2209 | 2405 |
| 24 | 411449.2 | g2154000 | 2210 | 2639 |
| 24 | 411449.2 | g3739242 | 2216 | 2561 |
| 24 | 411449.2 | 5276078H1 | 2220 | 2494 |
| 24 | 411449.2 | g889241 | 2223 | 2564 |
| 24 | 411449.2 | g1148058 | 2235 | 2554 |
| 24 | 411449.2 | g3077322 | 2268 | 2561 |
| 24 | 411449.2 | g784419 | 2295 | 2555 |
| 24 | 411449.2 | 2758260H1 | 2312 | 2541 |
| 24 | 411449.2 | g4109094 | 2349 | 2619 |
| 24 | 411449.2 | g1785173 | 2442 | 2554 |
| 24 | 411449.2 | 3504710H1 | 2444 | 2554 |
| 24 | 411449.2 | g3678027 | 2544 | 2899 |
| 24 | 411449.2 | g3412805 | 2545 | 2944 |
| 24 | 411449.2 | g3214749 | 2546 | 2950 |
| 24 | 411449.2 | g4149617 | 2546 | 2995 |
| 24 | 411449.2 | g2969810 | 2546 | 2703 |
| 24 | 411449.2 | g746614 | 2551 | 2916 |
| 24 | 411449.2 | g3756632 | 2550 | 2615 |
| 24 | 411449.2 | g2969693 | 2551 | 2970 |
| 24 | 411449.2 | g3770455 | 2552 | 2832 |
| 24 | 411449.2 | g2538262 | 2552 | 2618 |
| 24 | 411449.2 | g2882695 | 2552 | 2871 |
| 24 | 411449.2 | 3073355H1 | 3031 | 3337 |
| 24 | 411449.2 | 2502022H1 | 3034 | 3299 |
| 24 | 411449.2 | 2203860H1 | 3035 | 3321 |
| 24 | 411449.2 | 2415335H1 | 3040 | 3241 |
| 24 | 411449.2 | 4068029H1 | 3043 | 3337 |
| 24 | 411449.2 | 3633303H1 | 3044 | 3319 |
| 24 | 411449.2 | 1972353H1 | 3039 | 3301 |
| 24 | 411449.2 | g746725 | 3044 | 3326 |
| 24 | 411449.2 | 3737404H1 | 3044 | 3324 |
| 24 | 411449.2 | 3557887H1 | 3044 | 3317 |
| 24 | 411449.2 | 3295556H1 | 3046 | 3325 |
| 24 | 411449.2 | 4114476H1 | 3047 | 3337 |
| 24 | 411449.2 | 4795315H1 | 3048 | 3337 |
| 24 | 411449.2 | 577995H1 | 3047 | 3307 |
| 24 | 411449.2 | 3167827H1 | 3051 | 3339 |
| 24 | 411449.2 | 5686358H1 | 3055 | 3339 |
| 24 | 411449.2 | 2781948H1 | 3066 | 3331 |
| 24 | 411449.2 | 2561784H2 | 3068 | 3340 |
| 24 | 411449.2 | 4712882H1 | 3077 | 3339 |
| 24 | 411449.2 | 5374882H1 | 3086 | 3335 |
| 24 | 411449.2 | g2153894 | 3087 | 3334 |
| 24 | 411449.2 | 3596863H1 | 3133 | 3337 |
| 24 | 411449.2 | 4675071H1 | 3155 | 3312 |
| 24 | 411449.2 | 5285819H1 | 3161 | 3337 |
| 24 | 411449.2 | 3638749H1 | 3186 | 3301 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 24 | 411449.2 | 2707450H1 | 3198 | 3319 |
| 24 | 411449.2 | 5695240H1 | 3208 | 3337 |
| 24 | 411449.2 | 5059716H1 | 3251 | 3316 |
| 24 | 411449.2 | g1781709 | 1 | 493 |
| 24 | 411449.2 | g1775656 | 1 | 432 |
| 24 | 411449.2 | 2472307F6 | 18 | 397 |
| 24 | 411449.2 | 2472307H1 | 18 | 265 |
| 24 | 411449.2 | g822142 | 36 | 381 |
| 24 | 411449.2 | g2000779 | 51 | 258 |
| 24 | 411449.2 | 2782774H1 | 79 | 356 |
| 24 | 411449.2 | g1295851 | 138 | 810 |
| 24 | 411449.2 | g1312393 | 138 | 614 |
| 24 | 411449.2 | 3448516R6 | 142 | 624 |
| 24 | 411449.2 | g1010505 | 2552 | 2714 |
| 24 | 411449.2 | g1692011 | 2553 | 2956 |
| 24 | 411449.2 | 1274760F1 | 2555 | 2848 |
| 24 | 411449.2 | g2658624 | 2554 | 2974 |
| 24 | 411449.2 | g1921464 | 2554 | 2863 |
| 24 | 411449.2 | g1153184 | 2554 | 2764 |
| 24 | 411449.2 | g2768980 | 2555 | 2668 |
| 24 | 411449.2 | g2153787 | 2555 | 2972 |
| 24 | 411449.2 | g3117405 | 2555 | 2850 |
| 24 | 411449.2 | g2933847 | 2554 | 2995 |
| 24 | 411449.2 | g2934183 | 2555 | 3034 |
| 24 | 411449.2 | g2568950 | 2556 | 2868 |
| 24 | 411449.2 | 2018010H1 | 2562 | 2849 |
| 24 | 411449.2 | g1636435 | 2656 | 2796 |
| 24 | 411449.2 | 2050014H1 | 2689 | 2969 |
| 24 | 411449.2 | 2910217H1 | 2694 | 2967 |
| 24 | 411449.2 | g2153999 | 2709 | 3188 |
| 24 | 411449.2 | g2204848 | 2778 | 3188 |
| 24 | 411449.2 | 2476089H1 | 2798 | 3037 |
| 24 | 411449.2 | g1925326 | 2847 | 3323 |
| 24 | 411449.2 | g1678125 | 2913 | 3317 |
| 24 | 411449.2 | g2558304 | 2937 | 3101 |
| 24 | 411449.2 | g2100182 | 2951 | 3334 |
| 24 | 411449.2 | 506236H1 | 2971 | 3182 |
| 24 | 411449.2 | 3518329H1 | 2976 | 3311 |
| 24 | 411449.2 | 3473456H1 | 2979 | 3337 |
| 24 | 411449.2 | 1698871H1 | 2986 | 3230 |
| 24 | 411449.2 | 4512967H1 | 2987 | 3259 |
| 24 | 411449.2 | g1692106 | 2999 | 3198 |
| 24 | 411449.2 | 3081295H1 | 3009 | 3317 |
| 24 | 411449.2 | 4819058H1 | 3020 | 3324 |
| 24 | 411449.2 | 3492756H1 | 3028 | 3337 |
| 24 | 411449.2 | 3448616H1 | 142 | 413 |
| 24 | 411449.2 | 6138342H1 | 150 | 451 |
| 24 | 411449.2 | 1957671H1 | 154 | 438 |
| 24 | 411449.2 | g3888550 | 166 | 634 |
| 24 | 411449.2 | 239023H1 | 199 | 433 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 24 | 411449.2 | 463172H1 | 205 | 453 |
| 24 | 411449.2 | g1471862 | 257 | 732 |
| 24 | 411449.2 | g1471855 | 257 | 691 |
| 24 | 411449.2 | 5153884H1 | 352 | 608 |
| 24 | 411449.2 | 593640H1 | 385 | 596 |
| 24 | 411449.2 | g2037867 | 408 | 741 |
| 24 | 411449.2 | g2946071 | 438 | 791 |
| 25 | 18549.2 | 2666675H1 | 1 | 256 |
| 25 | 18549.2 | 2666675F6 | 1 | 447 |
| 25 | 18549.2 | 2872055H1 | 35 | 324 |
| 25 | 18549.2 | 3674803H1 | 45 | 340 |
| 25 | 18549.2 | g4393247 | 48 | 393 |
| 26 | 236043.3 | 4854788H1 | 57 | 322 |
| 26 | 236043.3 | 2637174H1 | 57 | 313 |
| 26 | 236043.3 | 4154039H1 | 57 | 326 |
| 26 | 236043.3 | 4154437H1 | 66 | 332 |
| 26 | 236043.3 | g1784274 | 72 | 327 |
| 26 | 236043.3 | g1782229 | 74 | 421 |
| 26 | 236043.3 | 5167751H1 | 1 | 259 |
| 26 | 236043.3 | 4154654H1 | 31 | 288 |
| 26 | 236043.3 | 2637174F6 | 57 | 407 |
| 26 | 236043.3 | 5856494H1 | 52 | 322 |
| 26 | 236043.3 | 753667H1 | 57 | 311 |
| 26 | 236043.3 | 2638855F6 | 615 | 912 |
| 26 | 236043.3 | 2638855T6 | 624 | 1270 |
| 26 | 236043.3 | 4000427H1 | 750 | 952 |
| 26 | 236043.3 | 118534H1 | 826 | 1086 |
| 26 | 236043.3 | g1782014 | 858 | 1286 |
| 26 | 236043.3 | g2553565 | 891 | 1313 |
| 26 | 236043.3 | g707651 | 915 | 1263 |
| 26 | 236043.3 | 506959H1 | 946 | 1251 |
| 26 | 236043.3 | g1792156 | 953 | 1283 |
| 26 | 236043.3 | g2616478 | 957 | 1266 |
| 26 | 236043.3 | g1792664 | 975 | 1281 |
| 26 | 236043.3 | g3754889 | 1007 | 1313 |
| 26 | 236043.3 | g707650 | 1011 | 1322 |
| 26 | 236043.3 | g1784007 | 1062 | 1313 |
| 26 | 236043.3 | g1782018 | 1086 | 1313 |
| 26 | 236043.3 | 1209542H1 | 1096 | 1326 |
| 26 | 236043.3 | 2433413H1 | 1134 | 1379 |
| 26 | 236043.3 | g1792600 | 1184 | 1283 |
| 26 | 236043.3 | 4401019T6 | 1210 | 1699 |
| 26 | 236043.3 | 2637174T6 | 1202 | 1260 |
| 26 | 236043.3 | g2342125 | 1329 | 1530 |
| 26 | 236043.3 | 118858H1 | 75 | 275 |
| 26 | 236043.3 | g1760400 | 77 | 392 |
| 26 | 236043.3 | g1664195 | 79 | 448 |
| 26 | 236043.3 | g1784031 | 82 | 542 |
| 26 | 236043.3 | 5298110H1 | 93 | 360 |
| 26 | 236043.3 | 5280294H1 | 110 | 358 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 26 | 236043.3 | 3516062H1 | 113 | 391 |
| 26 | 236043.3 | 4012792H1 | 114 | 402 |
| 26 | 236043.3 | 5856458H1 | 176 | 453 |
| 26 | 236043.3 | 5856757H1 | 176 | 310 |
| 26 | 236043.3 | g1784279 | 204 | 475 |
| 26 | 236043.3 | g707758 | 277 | 621 |
| 26 | 236043.3 | 4150892H1 | 309 | 572 |
| 26 | 236043.3 | g1792258 | 414 | 803 |
| 26 | 236043.3 | 3420770H1 | 432 | 622 |
| 26 | 236043.3 | 2638855H1 | 615 | 846 |
| 26 | 236043.3 | 121222H1 | 615 | 730 |
| 27 | 445433.2 | 3287379H1 | 1 | 244 |
| 27 | 445433.2 | 2275475H1 | 193 | 424 |
| 27 | 445433.2 | 3967290T6 | 308 | 705 |
| 27 | 445433.2 | 3967290F6 | 315 | 723 |
| 27 | 445433.2 | 4133307H1 | 315 | 574 |
| 27 | 445433.2 | 3967290H1 | 315 | 558 |
| 27 | 445433.2 | 1343762F6 | 681 | 1042 |
| 29 | 257121.2 | 2808987H1 | 805 | 1039 |
| 29 | 257121.2 | 4021166H1 | 831 | 1064 |
| 29 | 257121.2 | g1720349 | 2858 | 3353 |
| 29 | 257121.2 | g1522235 | 2858 | 3196 |
| 29 | 257121.2 | 3517880H1 | 2876 | 3155 |
| 29 | 257121.2 | 4459330H1 | 2876 | 3127 |
| 29 | 257121.2 | 3917168H1 | 2886 | 3150 |
| 29 | 257121.2 | 3160987H1 | 2897 | 3193 |
| 29 | 257121.2 | 4506429H1 | 2901 | 2981 |
| 29 | 257121.2 | 4624935H1 | 2909 | 3181 |
| 29 | 257121.2 | 879657H1 | 3382 | 3640 |
| 29 | 257121.2 | 434236H1 | 3388 | 3610 |
| 29 | 257121.2 | 2078370H1 | 2935 | 3207 |
| 29 | 257121.2 | 5058836H1 | 2958 | 3241 |
| 29 | 257121.2 | 1955941H1 | 2984 | 3250 |
| 29 | 257121.2 | 3993470H1 | 3015 | 3322 |
| 29 | 257121.2 | 2693754H1 | 3017 | 3296 |
| 29 | 257121.2 | 3528783H1 | 3020 | 3325 |
| 29 | 257121.2 | 2945721H1 | 3030 | 3196 |
| 29 | 257121.2 | 2947643H1 | 3028 | 3355 |
| 29 | 257121.2 | 5271360H1 | 3029 | 3276 |
| 29 | 257121.2 | g3840958 | 3443 | 3849 |
| 29 | 257121.2 | 406453H1 | 3445 | 3689 |
| 29 | 257121.2 | 865275H1 | 3451 | 3684 |
| 29 | 257121.2 | 2373951H1 | 3455 | 3695 |
| 29 | 257121.2 | g2329311 | 3465 | 3849 |
| 29 | 257121.2 | 5201726T6 | 3463 | 3823 |
| 29 | 257121.2 | 3929807H1 | 3466 | 3775 |
| 29 | 257121.2 | 4466537H1 | 3466 | 3723 |
| 29 | 257121.2 | 2071630H1 | 2781 | 3036 |
| 29 | 257121.2 | 2714703H1 | 2800 | 3056 |
| 29 | 257121.2 | 630218H1 | 2804 | 3049 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 29 | 257121.2 | g2219708 | 2818 | 3192 |
| 29 | 257121.2 | 5435973H1 | 2403 | 2648 |
| 29 | 257121.2 | g2189770 | 2409 | 2575 |
| 29 | 257121.2 | 3091740F6 | 2434 | 2892 |
| 29 | 257121.2 | 3091740H1 | 2435 | 2721 |
| 29 | 257121.2 | 4931375H1 | 2450 | 2729 |
| 29 | 257121.2 | 5086828H1 | 2466 | 2717 |
| 29 | 257121.2 | 4706266H1 | 2474 | 2701 |
| 29 | 257121.2 | 4623814H1 | 2489 | 2746 |
| 29 | 257121.2 | 1711729H1 | 2493 | 2682 |
| 29 | 257121.2 | 3808352H1 | 2493 | 2707 |
| 29 | 257121.2 | g1999011 | 2496 | 2841 |
| 29 | 257121.2 | 2080982H1 | 2500 | 2772 |
| 29 | 257121.2 | 4936350H1 | 2507 | 2800 |
| 29 | 257121.2 | 2473816H1 | 2534 | 2761 |
| 29 | 257121.2 | 2473816F6 | 2534 | 3077 |
| 29 | 257121.2 | 2228396H1 | 2540 | 2791 |
| 29 | 257121.2 | 4021166F6 | 831 | 1284 |
| 29 | 257121.2 | 2051255H1 | 838 | 1128 |
| 29 | 257121.2 | 3001820F6 | 870 | 1041 |
| 29 | 257121.2 | 3001820H1 | 871 | 1152 |
| 29 | 257121.2 | 5395159H1 | 935 | 1013 |
| 29 | 257121.2 | 492727H1 | 1100 | 1333 |
| 29 | 257121.2 | 2738108H1 | 1122 | 1378 |
| 29 | 257121.2 | 3344641H1 | 1132 | 1370 |
| 29 | 257121.2 | 526778H1 | 1241 | 1504 |
| 29 | 257121.2 | 310738H1 | 1296 | 1527 |
| 29 | 257121.2 | 2754594H1 | 1296 | 1556 |
| 29 | 257121.2 | 4055308H1 | 3034 | 3328 |
| 29 | 257121.2 | 4021166T6 | 3036 | 3558 |
| 29 | 257121.2 | 4465912H1 | 3059 | 3327 |
| 29 | 257121.2 | 5586546H1 | 3067 | 3312 |
| 29 | 257121.2 | g1239705 | 3068 | 3332 |
| 29 | 257121.2 | 552938H1 | 3096 | 3365 |
| 29 | 257121.2 | 3730966H1 | 3111 | 3436 |
| 29 | 257121.2 | g1444147 | 3121 | 3612 |
| 29 | 257121.2 | 5550566H1 | 3172 | 3413 |
| 29 | 257121.2 | 5507424H1 | 1840 | 2028 |
| 29 | 257121.2 | 5508405H1 | 1840 | 2087 |
| 29 | 257121.2 | 3000303H1 | 1849 | 2065 |
| 29 | 257121.2 | g1975045 | 1853 | 2120 |
| 29 | 257121.2 | 5662088H1 | 1854 | 2082 |
| 29 | 257121.2 | g3770956 | 1868 | 2224 |
| 29 | 257121.2 | g3802745 | 1899 | 2216 |
| 29 | 257121.2 | 4892961H1 | 1903 | 2189 |
| 29 | 257121.2 | 3479021H1 | 1924 | 2149 |
| 29 | 257121.2 | 1690528H1 | 1938 | 2158 |
| 29 | 257121.2 | 4467223H1 | 1953 | 2159 |
| 29 | 257121.2 | 5270694H1 | 3395 | 3651 |
| 29 | 257121.2 | g1719483 | 3568 | 3855 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 29 | 257121.2 | g4531585 | 3586 | 3849 |
| 29 | 257121.2 | 3720345H1 | 3594 | 3759 |
| 29 | 257121.2 | 2569980H1 | 3601 | 3849 |
| 29 | 257121.2 | g1203034 | 3610 | 3783 |
| 29 | 257121.2 | g4450994 | 3614 | 3849 |
| 29 | 257121.2 | 4901641H1 | 3614 | 3822 |
| 29 | 257121.2 | 4960977H1 | 2828 | 3096 |
| 29 | 257121.2 | g2728590 | 2834 | 3223 |
| 29 | 257121.2 | 3162492H1 | 2842 | 3125 |
| 29 | 257121.2 | g1719482 | 2858 | 3289 |
| 29 | 257121.2 | g2211339 | 3417 | 3848 |
| 29 | 257121.2 | g2659436 | 3422 | 3854 |
| 29 | 257121.2 | g4296041 | 3428 | 3849 |
| 29 | 257121.2 | 5921004H1 | 3429 | 3679 |
| 29 | 257121.2 | 2962704H1 | 3431 | 3741 |
| 29 | 257121.2 | 5922274H1 | 3431 | 3715 |
| 29 | 257121.2 | g2329245 | 3433 | 3703 |
| 29 | 257121.2 | g4328182 | 3435 | 3849 |
| 29 | 257121.2 | 1994936R6 | 3438 | 3848 |
| 29 | 257121.2 | 1994936T6 | 3438 | 3808 |
| 29 | 257121.2 | 1994936H1 | 3438 | 3705 |
| 29 | 257121.2 | 865275T1 | 3439 | 3794 |
| 29 | 257121.2 | 453069H1 | 1980 | 2189 |
| 29 | 257121.2 | 5982875H1 | 2023 | 2298 |
| 29 | 257121.2 | g2457731 | 2054 | 2450 |
| 29 | 257121.2 | 2883663H1 | 2077 | 2328 |
| 29 | 257121.2 | 1786163H1 | 2091 | 2273 |
| 29 | 257121.2 | 3433528H1 | 2097 | 2303 |
| 29 | 257121.2 | 2588747H2 | 2117 | 2375 |
| 29 | 257121.2 | 4715351H1 | 2157 | 2241 |
| 29 | 257121.2 | 4691328H1 | 2198 | 2462 |
| 29 | 257121.2 | 2904069H1 | 2226 | 2537 |
| 29 | 257121.2 | 4300967H1 | 2236 | 2508 |
| 29 | 257121.2 | 4624918H1 | 2909 | 3183 |
| 29 | 257121.2 | 2374252H1 | 2910 | 3120 |
| 29 | 257121.2 | 4909794H1 | 2922 | 3203 |
| 29 | 257121.2 | 1741105H1 | 2925 | 3159 |
| 29 | 257121.2 | 1741564H1 | 2925 | 3086 |
| 29 | 257121.2 | 1741564R6 | 2925 | 3384 |
| 29 | 257121.2 | g531805 | 3303 | 3701 |
| 29 | 257121.2 | 3091740T6 | 3306 | 3809 |
| 29 | 257121.2 | 2697055H1 | 3308 | 3597 |
| 29 | 257121.2 | 1672059T6 | 3344 | 3809 |
| 29 | 257121.2 | g2818402 | 3359 | 3858 |
| 29 | 257121.2 | 2423136H1 | 3377 | 3613 |
| 29 | 257121.2 | 879657T1 | 3382 | 3809 |
| 29 | 257121.2 | g4329746 | 3472 | 3849 |
| 29 | 257121.2 | g4243717 | 3477 | 3856 |
| 29 | 257121.2 | 3929886H1 | 3478 | 3759 |
| 29 | 257121.2 | g2619701 | 3487 | 3892 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 29 | 257121.2 | g2835679 | 3490 | 3849 |
| 29 | 257121.2 | g2216408 | 3495 | 3849 |
| 29 | 257121.2 | 3040078H1 | 3500 | 3780 |
| 29 | 257121.2 | g2261946 | 3498 | 3849 |
| 29 | 257121.2 | g1522116 | 3502 | 3855 |
| 29 | 257121.2 | g3190675 | 3510 | 3848 |
| 29 | 257121.2 | g3281587 | 3517 | 3852 |
| 29 | 257121.2 | g1186385 | 3517 | 3849 |
| 29 | 257121.2 | g3069539 | 3519 | 3856 |
| 29 | 257121.2 | g3229168 | 3523 | 3849 |
| 29 | 257121.2 | g2278678 | 3527 | 3855 |
| 29 | 257121.2 | g2595803 | 3528 | 3819 |
| 29 | 257121.2 | g1994549 | 3533 | 3860 |
| 29 | 257121.2 | g613905 | 3177 | 3513 |
| 29 | 257121.2 | 5905841H1 | 3186 | 3500 |
| 29 | 257121.2 | 1444396H1 | 3186 | 3461 |
| 29 | 257121.2 | g4187265 | 3195 | 3670 |
| 29 | 257121.2 | 2473816T6 | 3190 | 3808 |
| 29 | 257121.2 | g4186245 | 3195 | 3664 |
| 29 | 257121.2 | g3278632 | 3196 | 3616 |
| 29 | 257121.2 | g1506304 | 3213 | 3436 |
| 29 | 257121.2 | 1658291H1 | 3211 | 3453 |
| 29 | 257121.2 | g1994550 | 2249 | 2499 |
| 29 | 257121.2 | 3053733H1 | 2279 | 2581 |
| 29 | 257121.2 | 5641114H1 | 2298 | 2545 |
| 29 | 257121.2 | 995725R1 | 2300 | 2819 |
| 29 | 257121.2 | 995725H1 | 2300 | 2596 |
| 29 | 257121.2 | 2173193F6 | 2302 | 2653 |
| 29 | 257121.2 | 2173193H1 | 2302 | 2536 |
| 29 | 257121.2 | 3474978H1 | 2324 | 2601 |
| 29 | 257121.2 | 2106564H1 | 2329 | 2585 |
| 29 | 257121.2 | 5171463H1 | 2343 | 2629 |
| 29 | 257121.2 | 3449591H1 | 2355 | 2601 |
| 29 | 257121.2 | 2824470T6 | 2356 | 2941 |
| 29 | 257121.2 | 5546639H1 | 2387 | 2525 |
| 29 | 257121.2 | 5658068H1 | 2396 | 2642 |
| 29 | 257121.2 | 3771054H1 | 1386 | 1692 |
| 29 | 257121.2 | 2714558T6 | 1407 | 1749 |
| 29 | 257121.2 | g2037431 | 1419 | 1693 |
| 29 | 257121.2 | 2436316H1 | 1459 | 1650 |
| 29 | 257121.2 | g953893 | 1502 | 1633 |
| 29 | 257121.2 | 6075654H1 | 1487 | 1796 |
| 29 | 257121.2 | g1055687 | 1504 | 1601 |
| 29 | 257121.2 | g2034372 | 1557 | 1818 |
| 29 | 257121.2 | 1235661H1 | 1567 | 1836 |
| 29 | 257121.2 | 2824470F6 | 1573 | 1997 |
| 29 | 257121.2 | 2824470H1 | 1573 | 1809 |
| 29 | 257121.2 | g4509712 | 1585 | 1995 |
| 29 | 257121.2 | g3700825 | 1587 | 1927 |
| 29 | 257121.2 | 4876427H1 | 1618 | 1807 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 29 | 257121.2 | 3297894H1 | 1657 | 1904 |
| 29 | 257121.2 | g3764184 | 1663 | 1795 |
| 29 | 257121.2 | 4306167H1 | 1689 | 1809 |
| 29 | 257121.2 | 446319T6 | 1703 | 2175 |
| 29 | 257121.2 | 5984668H1 | 1802 | 2091 |
| 29 | 257121.2 | 5594456H1 | 1805 | 1931 |
| 29 | 257121.2 | 3202943H1 | 1818 | 2073 |
| 29 | 257121.2 | 309745H1 | 3396 | 3648 |
| 29 | 257121.2 | 4652103H1 | 3402 | 3556 |
| 29 | 257121.2 | g3058761 | 3407 | 3849 |
| 29 | 257121.2 | g3961167 | 3409 | 3844 |
| 29 | 257121.2 | g3178632 | 3413 | 3849 |
| 29 | 257121.2 | 5645258H1 | 190 | 462 |
| 29 | 257121.2 | g751765 | 209 | 449 |
| 29 | 257121.2 | g751766 | 344 | 636 |
| 29 | 257121.2 | 2478708H1 | 363 | 607 |
| 29 | 257121.2 | 5544313H1 | 370 | 582 |
| 29 | 257121.2 | 1911275H1 | 460 | 725 |
| 29 | 257121.2 | 1437425F1 | 511 | 985 |
| 29 | 257121.2 | 1437426H1 | 511 | 734 |
| 29 | 257121.2 | 1437425H1 | 511 | 736 |
| 29 | 257121.2 | 3697654H1 | 541 | 823 |
| 29 | 257121.2 | 3333937H1 | 589 | 769 |
| 29 | 257121.2 | 2639996H1 | 706 | 952 |
| 29 | 257121.2 | 446319R6 | 788 | 1254 |
| 29 | 257121.2 | 446319H1 | 788 | 1039 |
| 29 | 257121.2 | 3068429F7 | 1 | 419 |
| 29 | 257121.2 | 3068429H1 | 1 | 291 |
| 29 | 257121.2 | 2821269H1 | 126 | 301 |
| 29 | 257121.2 | 3160523H1 | 3217 | 3501 |
| 29 | 257121.2 | 5024976H1 | 3222 | 3513 |
| 29 | 257121.2 | 3894771H1 | 3241 | 3546 |
| 29 | 257121.2 | 3360915H1 | 3240 | 3354 |
| 29 | 257121.2 | 1427954T6 | 3250 | 3809 |
| 29 | 257121.2 | 1741564T6 | 3251 | 3811 |
| 29 | 257121.2 | 1892219H1 | 3250 | 3516 |
| 29 | 257121.2 | 217319T6 | 3255 | 3819 |
| 29 | 257121.2 | 1861145H1 | 3261 | 3586 |
| 29 | 257121.2 | 1531024H1 | 3769 | 3849 |
| 29 | 257121.2 | 2608136H1 | 3540 | 3797 |
| 29 | 257121.2 | 2608136T6 | 3533 | 3805 |
| 29 | 257121.2 | 2608136F6 | 3540 | 3853 |
| 29 | 257121.2 | g1210010 | 3540 | 3850 |
| 29 | 257121.2 | g2805071 | 3540 | 3827 |
| 29 | 257121.2 | g2218792 | 3542 | 3849 |
| 29 | 257121.2 | 2721666H1 | 3563 | 3814 |
| 29 | 257121.2 | 4901093H1 | 3614 | 3850 |
| 29 | 257121.2 | g3092920 | 3617 | 3818 |
| 29 | 257121.2 | 4347921H1 | 3623 | 3877 |
| 29 | 257121.2 | 3109370H1 | 3625 | 3723 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 29 | 257121.2 | g1757345 | 3644 | 3855 |
| 29 | 257121.2 | g1211126 | 3645 | 3857 |
| 29 | 257121.2 | g1505472 | 3647 | 3864 |
| 29 | 257121.2 | 5565971H1 | 3650 | 3853 |
| 29 | 257121.2 | 4613032H1 | 3668 | 3856 |
| 29 | 257121.2 | 892266H1 | 3672 | 3830 |
| 29 | 257121.2 | 3729288H1 | 3676 | 3853 |
| 29 | 257121.2 | g1813011 | 3677 | 3855 |
| 29 | 257121.2 | g2063050 | 3680 | 3849 |
| 29 | 257121.2 | 3729209T1 | 3681 | 3810 |
| 29 | 257121.2 | g2063735 | 3707 | 3849 |
| 29 | 257121.2 | g2882056 | 3741 | 3849 |
| 29 | 257121.2 | 1547677H1 | 3769 | 3844 |
| 29 | 257121.2 | 6024576H1 | 2558 | 2843 |
| 29 | 257121.2 | 4909221H1 | 2607 | 2901 |
| 29 | 257121.2 | 1427954H1 | 2634 | 2884 |
| 29 | 257121.2 | 5170995H1 | 2648 | 2874 |
| 29 | 257121.2 | 1427954F6 | 2665 | 3120 |
| 29 | 257121.2 | 946401H1 | 2666 | 2926 |
| 29 | 257121.2 | 1374579H1 | 2670 | 2923 |
| 29 | 257121.2 | 2261955H1 | 2684 | 2948 |
| 29 | 257121.2 | 3094766H1 | 2691 | 2994 |
| 29 | 257121.2 | 2741005H1 | 2693 | 2978 |
| 29 | 257121.2 | 4635067H1 | 2704 | 2978 |
| 29 | 257121.2 | 1007503H1 | 2718 | 3048 |
| 29 | 257121.2 | 5284328H1 | 2720 | 2957 |
| 29 | 257121.2 | 5064384H1 | 2731 | 2941 |
| 29 | 257121.2 | 1672059H1 | 2742 | 2924 |
| 29 | 257121.2 | 1672008H1 | 2742 | 2975 |
| 29 | 257121.2 | 1672059F6 | 2742 | 3123 |
| 29 | 257121.2 | 2410045H1 | 2774 | 3011 |
| 37 | 84399.1 | 2520472H1 | 1 | 226 |
| 37 | 84399.1 | g4148125 | 155 | 499 |
| 38 | 350044.1 | 3110061F7 | 1 | 276 |
| 38 | 350044.1 | 3110061H1 | 3 | 289 |
| 38 | 350044.1 | 4308349H1 | 158 | 426 |
| 38 | 350044.1 | 4308349F6 | 158 | 587 |
| 38 | 350044.1 | 5333549H1 | 185 | 413 |
| 38 | 350044.1 | 3399811H1 | 405 | 650 |
| 38 | 350044.1 | 2288313H1 | 496 | 629 |
| 38 | 350044.1 | 4637040H1 | 583 | 844 |
| 38 | 350044.1 | 4637040F6 | 582 | 916 |
| 38 | 350044.1 | 4308349T6 | 662 | 1020 |
| 38 | 350044.1 | 308559H1 | 664 | 884 |
| 38 | 350044.1 | g1803082 | 832 | 926 |
| 38 | 350044.1 | 3977826H1 | 852 | 968 |
| 39 | 441329.2 | g3751157 | 1 | 412 |
| 39 | 441329.2 | g1470664 | 342 | 578 |
| 39 | 441329.2 | g1395945 | 342 | 644 |
| 39 | 441329.2 | 4327736H1 | 355 | 622 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 39 | 441329.2 | 132848H1 | 360 | 540 |
| 39 | 441329.2 | 132849R6 | 360 | 805 |
| 39 | 441329.2 | 131890H1 | 360 | 563 |
| 39 | 441329.2 | 131890R6 | 360 | 802 |
| 39 | 441329.2 | g1165579 | 361 | 542 |
| 39 | 441329.2 | g1928191 | 371 | 650 |
| 39 | 441329.2 | 4204432F6 | 374 | 765 |
| 39 | 441329.2 | g1734964 | 378 | 763 |
| 39 | 441329.2 | 3357071H1 | 701 | 979 |
| 39 | 441329.2 | 3357071F6 | 701 | 1118 |
| 39 | 441329.2 | 131890T6 | 1077 | 1235 |
| 39 | 441329.2 | g1733361 | 1084 | 1235 |
| 39 | 441329.2 | 132849T6 | 1123 | 1235 |
| 39 | 441329.2 | 4204432T6 | 1128 | 1235 |
| 40 | 442401.2 | 3349655H1 | 1 | 327 |
| 40 | 442401.2 | 4309840H1 | 10 | 304 |
| 40 | 442401.2 | 4349106H1 | 25 | 238 |
| 40 | 442401.2 | 5043378H1 | 42 | 296 |
| 40 | 442401.2 | 4789236H1 | 44 | 123 |
| 40 | 442401.2 | 5320882H1 | 45 | 180 |
| 40 | 442401.2 | 2551237H1 | 64 | 319 |
| 40 | 442401.2 | 4664370H1 | 71 | 322 |
| 40 | 442401.2 | 3510753H1 | 75 | 389 |
| 40 | 442401.2 | 693783H1 | 80 | 286 |
| 40 | 442401.2 | 693783R6 | 83 | 554 |
| 40 | 442401.2 | 3865603H1 | 88 | 388 |
| 40 | 442401.2 | 2289862H1 | 86 | 326 |
| 40 | 442401.2 | 3681694H1 | 92 | 386 |
| 40 | 442401.2 | 693783T6 | 128 | 736 |
| 40 | 442401.2 | 519767H1 | 511 | 741 |
| 41 | 444933.2 | 3492265H1 | 7 | 310 |
| 41 | 444933.2 | g1501708 | 7 | 263 |
| 41 | 444933.2 | 3155609H1 | 7 | 99 |
| 41 | 444933.2 | g1474433 | 9 | 386 |
| 41 | 444933.2 | 3118539H1 | 9 | 310 |
| 41 | 444933.2 | 3295816H1 | 15 | 277 |
| 41 | 444933.2 | 1592931H1 | 29 | 229 |
| 41 | 444933.2 | 1592931F6 | 29 | 395 |
| 41 | 444933.2 | 1592931T6 | 43 | 579 |
| 41 | 444933.2 | 3436778H1 | 1 | 229 |
| 42 | 481129.4 | 999203H1 | 234 | 552 |
| 42 | 481129.4 | 999203T1 | 234 | 708 |
| 42 | 481129.4 | g3278054 | 236 | 751 |
| 42 | 481129.4 | 3868296H1 | 240 | 560 |
| 42 | 481129.4 | g4522433 | 240 | 752 |
| 42 | 481129.4 | 3137701H1 | 240 | 575 |
| 42 | 481129.4 | g3417930 | 241 | 746 |
| 42 | 481129.4 | 3868754H1 | 240 | 570 |
| 42 | 481129.4 | g4070375 | 243 | 752 |
| 42 | 481129.4 | 715742H1 | 248 | 618 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | 4950629H1 | 245 | 592 |
| 42 | 481129.4 | g3644668 | 250 | 755 |
| 42 | 481129.4 | 1577433H1 | 250 | 516 |
| 42 | 481129.4 | g3446453 | 265 | 747 |
| 42 | 481129.4 | 2055077H1 | 267 | 589 |
| 42 | 481129.4 | g2288119 | 272 | 746 |
| 42 | 481129.4 | 3358220H1 | 293 | 617 |
| 42 | 481129.4 | g3797800 | 295 | 747 |
| 42 | 481129.4 | g2113046 | 296 | 749 |
| 42 | 481129.4 | g2740428 | 301 | 747 |
| 42 | 481129.4 | g2268545 | 312 | 747 |
| 42 | 481129.4 | g4087282 | 314 | 750 |
| 42 | 481129.4 | g2017050 | 320 | 622 |
| 42 | 481129.4 | g1924476 | 320 | 647 |
| 42 | 481129.4 | g3735602 | 323 | 748 |
| 42 | 481129.4 | g4532931 | 324 | 747 |
| 42 | 481129.4 | g3734942 | 324 | 751 |
| 42 | 481129.4 | g2737174 | 325 | 747 |
| 42 | 481129.4 | g3960674 | 330 | 751 |
| 42 | 481129.4 | g2252043 | 328 | 746 |
| 42 | 481129.4 | g3330470 | 342 | 747 |
| 42 | 481129.4 | g3960381 | 342 | 755 |
| 42 | 481129.4 | g4452123 | 351 | 746 |
| 42 | 481129.4 | g3678389 | 352 | 747 |
| 42 | 481129.4 | 1632555H1 | 353 | 564 |
| 42 | 481129.4 | 1632539H1 | 353 | 571 |
| 42 | 481129.4 | 4228443H1 | 353 | 686 |
| 42 | 481129.4 | g3108753 | 369 | 749 |
| 42 | 481129.4 | g2319165 | 368 | 747 |
| 42 | 481129.4 | g3755458 | 374 | 754 |
| 42 | 481129.4 | g4086634 | 397 | 747 |
| 42 | 481129.4 | g2322475 | 404 | 746 |
| 42 | 481129.4 | g3750698 | 406 | 748 |
| 42 | 481129.4 | g2021853 | 411 | 746 |
| 42 | 481129.4 | 3099690H1 | 423 | 747 |
| 42 | 481129.4 | 4125491H1 | 425 | 735 |
| 42 | 481129.4 | 5107680H1 | 443 | 746 |
| 42 | 481129.4 | 1253080F1 | 454 | 746 |
| 42 | 481129.4 | 2741014H1 | 456 | 747 |
| 42 | 481129.4 | g4086632 | 472 | 747 |
| 42 | 481129.4 | 3381102H1 | 195 | 363 |
| 42 | 481129.4 | 3136744H1 | 132 | 344 |
| 42 | 481129.4 | 2485548H1 | 37 | 110 |
| 42 | 481129.4 | 3238530H1 | 36 | 320 |
| 42 | 481129.4 | 5332434H1 | 132 | 262 |
| 42 | 481129.4 | 1743373H1 | 37 | 342 |
| 42 | 481129.4 | 2398059H1 | 30 | 282 |
| 42 | 481129.4 | 5299970H1 | 36 | 291 |
| 42 | 481129.4 | 4857940H1 | 132 | 351 |
| 42 | 481129.4 | 982689H1 | 28 | 284 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | 1684768H1 | 34 | 296 |
| 42 | 481129.4 | 4901916H1 | 29 | 347 |
| 42 | 481129.4 | 4340959H1 | 25 | 317 |
| 42 | 481129.4 | 1275760H1 | 195 | 375 |
| 42 | 481129.4 | g1975121 | 132 | 336 |
| 42 | 481129.4 | 534025H1 | 30 | 348 |
| 42 | 481129.4 | 3555252H1 | 24 | 317 |
| 42 | 481129.4 | 1504151H1 | 30 | 329 |
| 42 | 481129.4 | 4652858H1 | 195 | 379 |
| 42 | 481129.4 | 4638612H1 | 132 | 320 |
| 42 | 481129.4 | 4801179H1 | 30 | 310 |
| 42 | 481129.4 | 1970489H1 | 35 | 306 |
| 42 | 481129.4 | 2497883H1 | 30 | 272 |
| 42 | 481129.4 | 2557615H1 | 132 | 324 |
| 42 | 481129.4 | 4079990H1 | 195 | 386 |
| 42 | 481129.4 | 3555091H1 | 31 | 350 |
| 42 | 481129.4 | g1635875 | 160 | 388 |
| 42 | 481129.4 | 3010470H1 | 132 | 368 |
| 42 | 481129.4 | g1685660 | 488 | 846 |
| 42 | 481129.4 | 2716085H1 | 492 | 747 |
| 42 | 481129.4 | g2669356 | 508 | 750 |
| 42 | 481129.4 | g2035379 | 508 | 768 |
| 42 | 481129.4 | g4330124 | 517 | 749 |
| 42 | 481129.4 | g1125228 | 503 | 768 |
| 42 | 481129.4 | g1237906 | 515 | 746 |
| 42 | 481129.4 | 1253080H1 | 547 | 746 |
| 42 | 481129.4 | 1924052H1 | 559 | 747 |
| 42 | 481129.4 | 1444584H1 | 557 | 747 |
| 42 | 481129.4 | 3089459H1 | 563 | 755 |
| 42 | 481129.4 | 1952233H1 | 596 | 746 |
| 42 | 481129.4 | g2269237 | 609 | 746 |
| 42 | 481129.4 | 2058183R6 | 620 | 747 |
| 42 | 481129.4 | 2058183H1 | 620 | 747 |
| 42 | 481129.4 | g3321468 | 627 | 746 |
| 42 | 481129.4 | g2277314 | 629 | 747 |
| 42 | 481129.4 | 4363320H1 | 652 | 742 |
| 42 | 481129.4 | g2752791 | 670 | 746 |
| 42 | 481129.4 | 3221190H1 | 49 | 104 |
| 42 | 481129.4 | 529956H1 | 197 | 303 |
| 42 | 481129.4 | 1781280H1 | 202 | 315 |
| 42 | 481129.4 | 586079H1 | 35 | 104 |
| 42 | 481129.4 | 644429H1 | 37 | 104 |
| 42 | 481129.4 | 4713384H1 | 32 | 97 |
| 42 | 481129.4 | 1434389H1 | 30 | 104 |
| 42 | 481129.4 | 2286414H1 | 30 | 104 |
| 42 | 481129.4 | 1648951H1 | 34 | 111 |
| 42 | 481129.4 | 3589355H1 | 36 | 111 |
| 42 | 481129.4 | 1916331H1 | 195 | 338 |
| 42 | 481129.4 | 880829H1 | 28 | 97 |
| 42 | 481129.4 | 2475170H1 | 195 | 350 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | g1301383 | 263 | 352 |
| 42 | 481129.4 | 4612347H1 | 169 | 315 |
| 42 | 481129.4 | 116571H1 | 25 | 83 |
| 42 | 481129.4 | g1023181 | 141 | 205 |
| 42 | 481129.4 | 835155H1 | 37 | 333 |
| 42 | 481129.4 | 4153116H1 | 132 | 329 |
| 42 | 481129.4 | 4300976H1 | 132 | 343 |
| 42 | 481129.4 | 962735H1 | 132 | 324 |
| 42 | 481129.4 | 4904613H2 | 132 | 339 |
| 42 | 481129.4 | 4904321H1 | 195 | 362 |
| 42 | 481129.4 | 4130260H1 | 132 | 334 |
| 42 | 481129.4 | 534025F1 | 1 | 184 |
| 42 | 481129.4 | 5165213H2 | 4 | 265 |
| 42 | 481129.4 | g3085628 | 1 | 187 |
| 42 | 481129.4 | g2836097 | 1 | 184 |
| 42 | 481129.4 | g2397633 | 1 | 179 |
| 42 | 481129.4 | g2669712 | 1 | 181 |
| 42 | 481129.4 | 811971T1 | 1 | 191 |
| 42 | 481129.4 | 811971H1 | 1 | 67 |
| 42 | 481129.4 | g1578417 | 1 | 184 |
| 42 | 481129.4 | g3016145 | 1 | 253 |
| 42 | 481129.4 | g1425361 | 1 | 232 |
| 42 | 481129.4 | g2237584 | 1 | 231 |
| 42 | 481129.4 | g4072976 | 1 | 234 |
| 42 | 481129.4 | g3076557 | 1 | 252 |
| 42 | 481129.4 | 3716227H1 | 1 | 125 |
| 42 | 481129.4 | g2350483 | 1 | 232 |
| 42 | 481129.4 | g1289724 | 1 | 184 |
| 42 | 481129.4 | 4830687H1 | 14 | 227 |
| 42 | 481129.4 | 1270745H1 | 1 | 82 |
| 42 | 481129.4 | g3423624 | 1 | 187 |
| 42 | 481129.4 | g1685572 | 1 | 232 |
| 42 | 481129.4 | g2907402 | 1 | 79 |
| 42 | 481129.4 | g2987865 | 1 | 235 |
| 42 | 481129.4 | 4664951H1 | 1 | 115 |
| 42 | 481129.4 | g1114304 | 1 | 162 |
| 42 | 481129.4 | 1694416H1 | 1 | 78 |
| 42 | 481129.4 | g4076242 | 1 | 234 |
| 42 | 481129.4 | g2806310 | 1 | 181 |
| 42 | 481129.4 | g3110520 | 1 | 231 |
| 42 | 481129.4 | g3178699 | 1 | 235 |
| 42 | 481129.4 | 5691870H1 | 1 | 56 |
| 42 | 481129.4 | g2408708 | 1 | 232 |
| 42 | 481129.4 | g3070820 | 1 | 231 |
| 42 | 481129.4 | g1194184 | 1 | 183 |
| 42 | 481129.4 | g3400227 | 1 | 296 |
| 42 | 481129.4 | g1080367 | 1 | 151 |
| 42 | 481129.4 | g2768589 | 1 | 231 |
| 42 | 481129.4 | g3804766 | 1 | 184 |
| 42 | 481129.4 | g3190855 | 1 | 232 |

TABLE 4

| SEQ ID NO: | T mplate ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | g1087452 | 1 | 267 |
| 42 | 481129.4 | 4065091H1 | 1 | 169 |
| 42 | 481129.4 | g2838959 | 1 | 236 |
| 42 | 481129.4 | 3091849H1 | 1 | 226 |
| 42 | 481129.4 | g2263870 | 1 | 183 |
| 42 | 481129.4 | g2279285 | 1 | 239 |
| 42 | 481129.4 | 2755370H1 | 1 | 184 |
| 42 | 481129.4 | 2757311H1 | 1 | 184 |
| 42 | 481129.4 | 863475H1 | 1 | 192 |
| 42 | 481129.4 | g1689421 | 1 | 183 |
| 42 | 481129.4 | 5528923H1 | 19 | 211 |
| 42 | 481129.4 | g1506021 | 1 | 184 |
| 42 | 481129.4 | 2467514H1 | 10 | 261 |
| 42 | 481129.4 | 2943819H1 | 15 | 312 |
| 42 | 481129.4 | g1043884 | 16 | 196 |
| 42 | 481129.4 | 4461552H1 | 5 | 231 |
| 42 | 481129.4 | g1281569 | 1 | 531 |
| 42 | 481129.4 | g727471 | 1 | 234 |
| 42 | 481129.4 | 4889929H1 | 29 | 314 |
| 42 | 481129.4 | 4985113H1 | 29 | 354 |
| 42 | 481129.4 | 1488272H1 | 5 | 319 |
| 42 | 481129.4 | 2450834F6 | 6 | 558 |
| 42 | 481129.4 | 2450834H1 | 6 | 287 |
| 42 | 481129.4 | g2027943 | 7 | 201 |
| 42 | 481129.4 | 2394311H2 | 8 | 123 |
| 42 | 481129.4 | 3372442H1 | 1 | 258 |
| 42 | 481129.4 | g3307272 | 28 | 238 |
| 42 | 481129.4 | 3773060H1 | 38 | 366 |
| 42 | 481129.4 | 4125777H1 | 10 | 329 |
| 42 | 481129.4 | 5159564H1 | 12 | 302 |
| 42 | 481129.4 | 3233460H1 | 12 | 297 |
| 42 | 481129.4 | 5552762H1 | 12 | 315 |
| 42 | 481129.4 | 593632H1 | 12 | 123 |
| 42 | 481129.4 | 4513470H1 | 13 | 313 |
| 42 | 481129.4 | 5115162H1 | 44 | 356 |
| 42 | 481129.4 | 5039159H2 | 15 | 286 |
| 42 | 481129.4 | 2134140H1 | 15 | 328 |
| 42 | 481129.4 | 5943202H1 | 44 | 374 |
| 42 | 481129.4 | 027622R6 | 17 | 589 |
| 42 | 481129.4 | 027622H1 | 17 | 197 |
| 42 | 481129.4 | 3115953H1 | 46 | 372 |
| 42 | 481129.4 | 3489959H1 | 46 | 382 |
| 42 | 481129.4 | 5166761H1 | 44 | 322 |
| 42 | 481129.4 | 3626341H1 | 18 | 386 |
| 42 | 481129.4 | 3645933H1 | 48 | 327 |
| 42 | 481129.4 | 3159541H1 | 19 | 348 |
| 42 | 481129.4 | 2741058H1 | 52 | 282 |
| 42 | 481129.4 | 1475821H1 | 50 | 252 |
| 42 | 481129.4 | 4567661H1 | 22 | 350 |
| 42 | 481129.4 | 3601462H1 | 22 | 394 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | 4595577H1 | 53 | 283 |
| 42 | 481129.4 | 3819717H1 | 22 | 426 |
| 42 | 481129.4 | 4975524H1 | 22 | 359 |
| 42 | 481129.4 | 2084168H1 | 50 | 348 |
| 42 | 481129.4 | 5529123H1 | 18 | 212 |
| 42 | 481129.4 | 5700770H1 | 52 | 344 |
| 42 | 481129.4 | 5701471H1 | 52 | 355 |
| 42 | 481129.4 | g3134982 | 45 | 234 |
| 42 | 481129.4 | 4979603H1 | 52 | 355 |
| 42 | 481129.4 | 3202264H1 | 23 | 321 |
| 42 | 481129.4 | 4938763H1 | 16 | 130 |
| 42 | 481129.4 | 4202906H1 | 8 | 274 |
| 42 | 481129.4 | 526404H1 | 22 | 322 |
| 42 | 481129.4 | 2665708H1 | 21 | 280 |
| 42 | 481129.4 | 2941705H1 | 22 | 327 |
| 42 | 481129.4 | g1321143 | 11 | 346 |
| 42 | 481129.4 | g2021854 | 22 | 337 |
| 42 | 481129.4 | 1833977H1 | 22 | 134 |
| 42 | 481129.4 | 2100632H1 | 12 | 285 |
| 42 | 481129.4 | 3675543H1 | 12 | 320 |
| 42 | 481129.4 | 3186510H1 | 13 | 400 |
| 42 | 481129.4 | 3229333H1 | 14 | 321 |
| 42 | 481129.4 | 4885571H1 | 19 | 306 |
| 42 | 481129.4 | 3074685H1 | 16 | 321 |
| 42 | 481129.4 | 3663968H1 | 16 | 329 |
| 42 | 481129.4 | g2154555 | 1 | 29 |
| 42 | 481129.4 | 863475R1 | 1 | 192 |
| 42 | 481129.4 | 4230048H1 | 1 | 176 |
| 42 | 481129.4 | 3693558H1 | 18 | 328 |
| 42 | 481129.4 | g1648658 | 1 | 23 |
| 42 | 481129.4 | 584205H1 | 18 | 309 |
| 42 | 481129.4 | 4227248H1 | 21 | 324 |
| 42 | 481129.4 | g1043979 | 28 | 415 |
| 42 | 481129.4 | g2142101 | 1 | 24 |
| 42 | 481129.4 | 4203722H1 | 22 | 321 |
| 42 | 481129.4 | 4174825H1 | 22 | 337 |
| 42 | 481129.4 | 5115519H1 | 23 | 321 |
| 42 | 481129.4 | 3133373H1 | 25 | 334 |
| 42 | 481129.4 | 1275760F6 | 1 | 147 |
| 42 | 481129.4 | 1275760F1 | 1 | 76 |
| 42 | 481129.4 | 2051350H1 | 27 | 350 |
| 42 | 481129.4 | 4341005H1 | 27 | 398 |
| 42 | 481129.4 | 3134845H1 | 1 | 260 |
| 42 | 481129.4 | 3484150H1 | 1 | 316 |
| 42 | 481129.4 | 4584348H1 | 1 | 279 |
| 42 | 481129.4 | g2004902 | 1 | 283 |
| 42 | 481129.4 | 4374817H1 | 1 | 288 |
| 42 | 481129.4 | g1967486 | 1 | 183 |
| 42 | 481129.4 | 4147580H1 | 1 | 267 |
| 42 | 481129.4 | g1425470 | 1 | 78 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | 4276424H1 | 1 | 301 |
| 42 | 481129.4 | 082654H1 | 15 | 171 |
| 42 | 481129.4 | g3003216 | 1 | 306 |
| 42 | 481129.4 | g1648659 | 1 | 183 |
| 42 | 481129.4 | 4563261H1 | 1 | 255 |
| 42 | 481129.4 | 4563351H1 | 12 | 136 |
| 42 | 481129.4 | 116571F1 | 1 | 233 |
| 42 | 481129.4 | g4372584 | 1 | 234 |
| 42 | 481129.4 | 3994691H1 | 23 | 288 |
| 42 | 481129.4 | g2437312 | 1 | 188 |
| 42 | 481129.4 | 136145F1 | 1 | 184 |
| 42 | 481129.4 | 2768229H1 | 1 | 256 |
| 42 | 481129.4 | 5852546H1 | 1 | 272 |
| 42 | 481129.4 | 5844492H1 | 6 | 279 |
| 42 | 481129.4 | 1727158H1 | 8 | 249 |
| 42 | 481129.4 | 1950091H1 | 8 | 264 |
| 42 | 481129.4 | 3204976H1 | 10 | 290 |
| 42 | 481129.4 | 4554858H1 | 10 | 275 |
| 42 | 481129.4 | 5840991H2 | 11 | 303 |
| 42 | 481129.4 | 5840890H2 | 11 | 314 |
| 42 | 481129.4 | 4559105H1 | 12 | 298 |
| 42 | 481129.4 | 871105H1 | 12 | 123 |
| 42 | 481129.4 | 4998348H1 | 12 | 293 |
| 42 | 481129.4 | 871105R1 | 12 | 123 |
| 42 | 481129.4 | 4770486H1 | 13 | 287 |
| 42 | 481129.4 | 3426271H1 | 14 | 295 |
| 42 | 481129.4 | 3273645H1 | 14 | 298 |
| 42 | 481129.4 | 4502129H1 | 14 | 202 |
| 42 | 481129.4 | 3403224H1 | 15 | 221 |
| 42 | 481129.4 | 1511018H1 | 15 | 89 |
| 42 | 481129.4 | 4502229H1 | 15 | 288 |
| 42 | 481129.4 | 4042436H1 | 8 | 283 |
| 42 | 481129.4 | 3845464H1 | 10 | 357 |
| 42 | 481129.4 | 3720453H1 | 12 | 340 |
| 42 | 481129.4 | 1998071H1 | 22 | 123 |
| 42 | 481129.4 | 5835243H1 | 22 | 290 |
| 42 | 481129.4 | 3801167H1 | 22 | 236 |
| 42 | 481129.4 | 746353H1 | 12 | 243 |
| 42 | 481129.4 | g1685558 | 13 | 441 |
| 42 | 481129.4 | 2537659H1 | 15 | 303 |
| 42 | 481129.4 | 1728205H1 | 13 | 209 |
| 42 | 481129.4 | 2674637H1 | 15 | 271 |
| 42 | 481129.4 | 4855434H1 | 14 | 293 |
| 42 | 481129.4 | 2943583H2 | 15 | 322 |
| 42 | 481129.4 | 116571R1 | 1 | 182 |
| 42 | 481129.4 | g1320320 | 1 | 363 |
| 42 | 481129.4 | 2801342H1 | 5 | 264 |
| 42 | 481129.4 | 5528955H1 | 21 | 208 |
| 42 | 481129.4 | 3618142H1 | 24 | 377 |
| 42 | 481129.4 | 122118H1 | 17 | 164 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | 3580438H1 | 69 | 320 |
| 42 | 481129.4 | g1638276 | 25 | 435 |
| 42 | 481129.4 | 833841T1 | 40 | 708 |
| 42 | 481129.4 | 3997421H1 | 26 | 345 |
| 42 | 481129.4 | 4110340H1 | 56 | 374 |
| 42 | 481129.4 | 2403135H1 | 58 | 320 |
| 42 | 481129.4 | g1969685 | 58 | 472 |
| 42 | 481129.4 | 3026745H1 | 57 | 366 |
| 42 | 481129.4 | 833841H1 | 40 | 406 |
| 42 | 481129.4 | 5681519H1 | 42 | 198 |
| 42 | 481129.4 | 3987429H1 | 58 | 376 |
| 42 | 481129.4 | 5177261H2 | 70 | 330 |
| 42 | 481129.4 | 880829R1 | 29 | 768 |
| 42 | 481129.4 | g1954123 | 46 | 346 |
| 42 | 481129.4 | 755972H1 | 59 | 323 |
| 42 | 481129.4 | 755972R1 | 59 | 709 |
| 42 | 481129.4 | 3944829H1 | 77 | 366 |
| 42 | 481129.4 | 1546083H1 | 58 | 262 |
| 42 | 481129.4 | g1496701 | 1 | 184 |
| 42 | 481129.4 | 962735R2 | 78 | 716 |
| 42 | 481129.4 | 2561965H1 | 32 | 359 |
| 42 | 481129.4 | g2008330 | 80 | 340 |
| 42 | 481129.4 | 881179T1 | 30 | 687 |
| 42 | 481129.4 | 534025R1 | 32 | 572 |
| 42 | 481129.4 | 871105T1 | 54 | 123 |
| 42 | 481129.4 | g1956736 | 64 | 347 |
| 42 | 481129.4 | 4800988H1 | 83 | 224 |
| 42 | 481129.4 | g1685765 | 83 | 403 |
| 42 | 481129.4 | 1230023H1 | 32 | 172 |
| 42 | 481129.4 | 1805729H1 | 27 | 189 |
| 42 | 481129.4 | 1320892H1 | 38 | 321 |
| 42 | 481129.4 | 1599902H1 | 38 | 290 |
| 42 | 481129.4 | 4588734H1 | 85 | 374 |
| 42 | 481129.4 | g1955622 | 70 | 445 |
| 42 | 481129.4 | 3517903H1 | 40 | 255 |
| 42 | 481129.4 | 136145R1 | 69 | 590 |
| 42 | 481129.4 | 4979660H1 | 40 | 354 |
| 42 | 481129.4 | 1741320H1 | 42 | 338 |
| 42 | 481129.4 | 5117019H1 | 88 | 386 |
| 42 | 481129.4 | g1924567 | 88 | 536 |
| 42 | 481129.4 | 2343536H1 | 89 | 355 |
| 42 | 481129.4 | 4536803H1 | 90 | 398 |
| 42 | 481129.4 | g2158823 | 96 | 655 |
| 42 | 481129.4 | g1958858 | 97 | 676 |
| 42 | 481129.4 | g828123 | 93 | 501 |
| 42 | 481129.4 | g3988556 | 97 | 621 |
| 42 | 481129.4 | g3096874 | 99 | 539 |
| 42 | 481129.4 | g3178165 | 107 | 566 |
| 42 | 481129.4 | g4079967 | 110 | 541 |
| 42 | 481129.4 | 465693H1 | 66 | 199 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 42 | 481129.4 | g3003210 | 111 | 467 |
| 42 | 481129.4 | 5265746H1 | 120 | 374 |
| 42 | 481129.4 | 966316H1 | 118 | 374 |
| 42 | 481129.4 | 1458085H1 | 122 | 396 |
| 42 | 481129.4 | 1457484H1 | 122 | 305 |
| 42 | 481129.4 | 2109149H1 | 122 | 442 |
| 42 | 481129.4 | 2109149R6 | 123 | 704 |
| 42 | 481129.4 | g1243536 | 77 | 215 |
| 42 | 481129.4 | 958148H1 | 122 | 431 |
| 42 | 481129.4 | 5020591T1 | 124 | 705 |
| 42 | 481129.4 | g1087561 | 133 | 459 |
| 42 | 481129.4 | g1958920 | 135 | 700 |
| 42 | 481129.4 | g1147415 | 139 | 444 |
| 42 | 481129.4 | g3890878 | 103 | 203 |
| 42 | 481129.4 | 2401220H1 | 144 | 390 |
| 42 | 481129.4 | 3223013H1 | 149 | 494 |
| 42 | 481129.4 | 1680890H1 | 149 | 361 |
| 42 | 481129.4 | g1298313 | 155 | 667 |
| 42 | 481129.4 | 828345H1 | 143 | 226 |
| 42 | 481129.4 | 3441046H1 | 159 | 442 |
| 42 | 481129.4 | 2109149T6 | 163 | 707 |
| 42 | 481129.4 | 2004957H1 | 176 | 484 |
| 42 | 481129.4 | 6063286H1 | 180 | 511 |
| 42 | 481129.4 | 5020591H1 | 179 | 501 |
| 42 | 481129.4 | 2450834T6 | 182 | 707 |
| 42 | 481129.4 | 027622T6 | 183 | 830 |
| 42 | 481129.4 | 5597771H1 | 189 | 470 |
| 42 | 481129.4 | 821831T6 | 192 | 705 |
| 42 | 481129.4 | 821831R1 | 192 | 746 |
| 42 | 481129.4 | 821831R6 | 192 | 746 |
| 42 | 481129.4 | 821831F1 | 192 | 757 |
| 42 | 481129.4 | g2252207 | 199 | 748 |
| 42 | 481129.4 | 1457484R1 | 199 | 746 |
| 42 | 481129.4 | g3736470 | 203 | 746 |
| 42 | 481129.4 | g1295465 | 204 | 757 |
| 42 | 481129.4 | 5185783H1 | 205 | 496 |
| 42 | 481129.4 | g2115256 | 211 | 352 |
| 42 | 481129.4 | 4085165H1 | 215 | 541 |
| 42 | 481129.4 | g1577977 | 221 | 716 |
| 42 | 481129.4 | g3593921 | 222 | 750 |
| 42 | 481129.4 | g3429019 | 222 | 747 |
| 42 | 481129.4 | 5221813H2 | 225 | 517 |
| 42 | 481129.4 | 5863560H1 | 232 | 539 |
| 42 | 481129.4 | 999208H1 | 234 | 544 |
| 42 | 481129.4 | 999203R1 | 234 | 750 |
| 42 | 481129.4 | g3649202 | 235 | 753 |
| 42 | 481129.4 | g2265393 | 234 | 751 |
| 42 | 481129.4 | 4049323H1 | 235 | 584 |
| 43 | 481999.1 | 1255239H1 | 838 | 1104 |
| 43 | 481999.1 | 4913234H1 | 397 | 512 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 43 | 481999.1 | 1724376H1 | 864 | 932 |
| 43 | 481999.1 | 1509213H1 | 819 | 1012 |
| 43 | 481999.1 | 1849705H1 | 281 | 501 |
| 43 | 481999.1 | 1509213F6 | 819 | 1104 |
| 43 | 481999.1 | 1476972F6 | 260 | 574 |
| 43 | 481999.1 | g2195301 | 866 | 1104 |
| 43 | 481999.1 | g985997 | 864 | 941 |
| 43 | 481999.1 | 880602R1 | 397 | 512 |
| 43 | 481999.1 | g1980533 | 1 | 283 |
| 43 | 481999.1 | g2195281 | 987 | 1104 |
| 43 | 481999.1 | 5526658H1 | 780 | 991 |
| 43 | 481999.1 | 880602H1 | 394 | 457 |
| 43 | 481999.1 | 1255206H1 | 838 | 1104 |
| 43 | 481999.1 | 1509479F6 | 819 | 1104 |
| 43 | 481999.1 | 3457008H1 | 397 | 481 |
| 43 | 481999.1 | 3641330H1 | 723 | 1014 |
| 43 | 481999.1 | 4181741H1 | 717 | 799 |
| 43 | 481999.1 | 3384087F6 | 604 | 857 |
| 43 | 481999.1 | 4760678H1 | 1 | 286 |
| 43 | 481999.1 | 1509221H1 | 819 | 1006 |
| 43 | 481999.1 | 1509455H1 | 819 | 997 |
| 43 | 481999.1 | 2603491H1 | 880 | 1104 |
| 43 | 481999.1 | 170070H1 | 797 | 982 |
| 43 | 481999.1 | 3427420F6 | 521 | 604 |
| 43 | 481999.1 | 4760678F6 | 1 | 499 |
| 43 | 481999.1 | 4913234F6 | 235 | 670 |
| 43 | 481999.1 | 4072789H1 | 237 | 499 |
| 43 | 481999.1 | 1524891F6 | 269 | 890 |
| 43 | 481999.1 | 398973R1 | 305 | 830 |
| 43 | 481999.1 | 4972169H1 | 420 | 695 |
| 43 | 481999.1 | 4357304H1 | 435 | 710 |
| 43 | 481999.1 | 4181741F6 | 718 | 1107 |
| 43 | 481999.1 | 3123901F6 | 380 | 434 |
| 43 | 481999.1 | 5585642H1 | 650 | 868 |
| 43 | 481999.1 | 1724376F6 | 864 | 941 |
| 43 | 481999.1 | 2938884H1 | 864 | 941 |
| 43 | 481999.1 | 3369038H1 | 429 | 706 |
| 43 | 481999.1 | g1148977 | 1 | 387 |
| 43 | 481999.1 | g1971119 | 842 | 1104 |
| 43 | 481999.1 | g1761297 | 704 | 1005 |
| 43 | 481999.1 | 1476972H1 | 260 | 455 |
| 43 | 481999.1 | 4211180H1 | 642 | 873 |
| 43 | 481999.1 | g3665013 | 371 | 434 |
| 43 | 481999.1 | 5153580H1 | 813 | 1066 |
| 43 | 481999.1 | g2141617 | 440 | 513 |
| 43 | 481999.1 | 1509479H1 | 819 | 997 |
| 46 | 338992.1 | 5510632H1 | 1 | 202 |
| 46 | 338992.1 | 4782751F6 | 127 | 187 |
| 46 | 338992.1 | 4782751H1 | 127 | 372 |
| 46 | 338992.1 | 2987413H1 | 192 | 478 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 46 | 338992.1 | 6064208H1 | 204 | 465 |
| 46 | 338992.1 | g1984325 | 272 | 505 |
| 46 | 338992.1 | g3239921 | 294 | 653 |
| 46 | 338992.1 | g2026718 | 300 | 591 |
| 46 | 338992.1 | g2816869 | 321 | 569 |
| 46 | 338992.1 | 4760384H1 | 492 | 580 |
| 46 | 338992.1 | 4760384F6 | 492 | 965 |
| 46 | 338992.1 | 1305496F6 | 802 | 1270 |
| 46 | 338992.1 | 1305496H1 | 802 | 945 |
| 46 | 338992.1 | 3617505H1 | 823 | 1024 |
| 46 | 338992.1 | g875546 | 878 | 1132 |
| 46 | 338992.1 | g669800 | 877 | 1199 |
| 46 | 338992.1 | g874649 | 878 | 1180 |
| 46 | 338992.1 | g771213 | 1011 | 1062 |
| 46 | 338992.1 | 3211592H1 | 1056 | 1292 |
| 46 | 338992.1 | 093259H1 | 1081 | 1322 |
| 46 | 338992.1 | 926880H1 | 1123 | 1363 |
| 46 | 338992.1 | 1366443H1 | 1137 | 1385 |
| 46 | 338992.1 | 1366443R6 | 1137 | 1466 |
| 46 | 338992.1 | 4960778H1 | 1160 | 1431 |
| 46 | 338992.1 | 4625867H1 | 1224 | 1487 |
| 46 | 338992.1 | 3383444H1 | 1259 | 1504 |
| 46 | 338992.1 | 5640688H1 | 1417 | 1657 |
| 46 | 338992.1 | 4030718H1 | 1568 | 1830 |
| 46 | 338992.1 | 4030718F6 | 1568 | 1945 |
| 46 | 338992.1 | 2742284H1 | 1620 | 1870 |
| 51 | 206603.1 | g3278540 | 1 | 465 |
| 51 | 206603.1 | g4390504 | 25 | 492 |
| 51 | 206603.1 | g3870096 | 41 | 504 |
| 51 | 206603.1 | g1109407 | 277 | 604 |
| 51 | 206603.1 | 5631413H1 | 367 | 608 |
| 51 | 206603.1 | 5631413F6 | 367 | 812 |
| 51 | 206603.1 | 1992082H1 | 531 | 726 |
| 51 | 206603.1 | 6092061H1 | 626 | 906 |
| 52 | 435694.2 | 4936540H1 | 1112 | 1364 |
| 52 | 435694.2 | 5021371H1 | 1139 | 1417 |
| 52 | 435694.2 | 3726226H1 | 1155 | 1450 |
| 52 | 435694.2 | g2015121 | 1170 | 1366 |
| 52 | 435694.2 | 456419H1 | 1185 | 1425 |
| 52 | 435694.2 | 461227R6 | 1185 | 1685 |
| 52 | 435694.2 | 456301H1 | 1185 | 1433 |
| 52 | 435694.2 | 461227H1 | 1185 | 1446 |
| 52 | 435694.2 | 460920H1 | 1185 | 1432 |
| 52 | 435694.2 | 461106H1 | 1185 | 1440 |
| 52 | 435694.2 | 457987H1 | 1185 | 1424 |
| 52 | 435694.2 | 458221H1 | 1185 | 1435 |
| 52 | 435694.2 | 454834R1 | 1185 | 1705 |
| 52 | 435694.2 | 461049H1 | 1185 | 1425 |
| 52 | 435694.2 | 454834H1 | 1185 | 1424 |
| 52 | 435694.2 | 4957324H1 | 266 | 486 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 52 | 435694.2 | 3390485H1 | 267 | 486 |
| 52 | 435694.2 | 1437407H1 | 243 | 486 |
| 52 | 435694.2 | 5902323H1 | 243 | 486 |
| 52 | 435694.2 | 3676724H1 | 271 | 486 |
| 52 | 435694.2 | 1437407F6 | 243 | 486 |
| 52 | 435694.2 | 1436837H1 | 243 | 480 |
| 52 | 435694.2 | 5900454H1 | 243 | 486 |
| 52 | 435694.2 | 2706491H1 | 270 | 566 |
| 52 | 435694.2 | 2837043H1 | 244 | 486 |
| 52 | 435694.2 | 3165547H1 | 246 | 486 |
| 52 | 435694.2 | 5903233H1 | 247 | 568 |
| 52 | 435694.2 | 3687630H1 | 251 | 565 |
| 52 | 435694.2 | 4268779H1 | 252 | 486 |
| 52 | 435694.2 | 3460048H1 | 251 | 487 |
| 52 | 435694.2 | 3077826H1 | 253 | 501 |
| 52 | 435694.2 | 4121058H1 | 271 | 486 |
| 52 | 435694.2 | 4900932H1 | 273 | 562 |
| 52 | 435694.2 | 3615215F6 | 280 | 731 |
| 52 | 435694.2 | 832322H1 | 252 | 486 |
| 52 | 435694.2 | 3615215H1 | 280 | 588 |
| 52 | 435694.2 | 2406515H1 | 254 | 486 |
| 52 | 435694.2 | 3141771H1 | 327 | 608 |
| 52 | 435694.2 | 3615215T6 | 351 | 720 |
| 52 | 435694.2 | g1774842 | 351 | 728 |
| 52 | 435694.2 | 3932194H1 | 351 | 486 |
| 52 | 435694.2 | 4003539H1 | 351 | 589 |
| 52 | 435694.2 | g587196 | 351 | 593 |
| 52 | 435694.2 | 1285417F6 | 400 | 708 |
| 52 | 435694.2 | 1285417H1 | 400 | 637 |
| 52 | 435694.2 | 2763950H1 | 502 | 762 |
| 52 | 435694.2 | 2415062H1 | 512 | 710 |
| 52 | 435694.2 | 2415070H1 | 512 | 702 |
| 52 | 435694.2 | 3860483H1 | 517 | 813 |
| 52 | 435694.2 | g1766458 | 520 | 850 |
| 52 | 435694.2 | g3076400 | 519 | 940 |
| 52 | 435694.2 | 823127H1 | 556 | 831 |
| 52 | 435694.2 | g3190613 | 562 | 745 |
| 52 | 435694.2 | 2926224H2 | 688 | 999 |
| 52 | 435694.2 | 2708490H1 | 690 | 942 |
| 52 | 435694.2 | g4311248 | 702 | 1175 |
| 52 | 435694.2 | g4389608 | 707 | 1164 |
| 52 | 435694.2 | g3700432 | 738 | 1174 |
| 52 | 435694.2 | g1773867 | 738 | 1172 |
| 52 | 435694.2 | 1285417T6 | 758 | 904 |
| 52 | 435694.2 | g4111706 | 797 | 1172 |
| 52 | 435694.2 | 774013H1 | 825 | 1036 |
| 52 | 435694.2 | 1454843H1 | 899 | 1163 |
| 52 | 435694.2 | 1450887F1 | 899 | 1172 |
| 52 | 435694.2 | 3742523H1 | 896 | 1163 |
| 52 | 435694.2 | 1450887H1 | 899 | 1152 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 52 | 435694.2 | 3742566H1 | 899 | 1163 |
| 52 | 435694.2 | 1454843T6 | 913 | 1128 |
| 52 | 435694.2 | 4710458H1 | 961 | 1249 |
| 52 | 435694.2 | 4977443H1 | 1005 | 1268 |
| 52 | 435694.2 | 3237593H1 | 1060 | 1304 |
| 52 | 435694.2 | 4705379T6 | 1071 | 1144 |
| 52 | 435694.2 | 4853112H1 | 1100 | 1295 |
| 52 | 435694.2 | 2666646T6 | 1628 | 2175 |
| 52 | 435694.2 | 2959842H1 | 1635 | 1861 |
| 52 | 435694.2 | 2963934T6 | 1633 | 2162 |
| 52 | 435694.2 | 3639070H1 | 1655 | 1967 |
| 52 | 435694.2 | 4322530H1 | 1660 | 1782 |
| 52 | 435694.2 | 5062716H1 | 1673 | 1795 |
| 52 | 435694.2 | 1710033H1 | 1703 | 1959 |
| 52 | 435694.2 | 5021371T1 | 1715 | 2162 |
| 52 | 435694.2 | 3894718H1 | 1752 | 1968 |
| 52 | 435694.2 | 2151922H1 | 1765 | 2018 |
| 52 | 435694.2 | g1645818 | 1810 | 2162 |
| 52 | 435694.2 | 2323908H1 | 1816 | 1878 |
| 52 | 435694.2 | g4070415 | 1817 | 2162 |
| 52 | 435694.2 | g4175351 | 1821 | 2162 |
| 52 | 435694.2 | g3595864 | 1824 | 2162 |
| 52 | 435694.2 | g3434496 | 1843 | 2162 |
| 52 | 435694.2 | g1775010 | 1845 | 2162 |
| 52 | 435694.2 | g1645819 | 1872 | 2162 |
| 52 | 435694.2 | g1164565 | 1885 | 2262 |
| 52 | 435694.2 | 4701327H1 | 1906 | 2173 |
| 52 | 435694.2 | g1766363 | 1910 | 2162 |
| 52 | 435694.2 | 2253564R6 | 1939 | 2171 |
| 52 | 435694.2 | 2253564H1 | 1939 | 2195 |
| 52 | 435694.2 | 4540530H1 | 1951 | 2181 |
| 52 | 435694.2 | 4540144H1 | 1953 | 2199 |
| 52 | 435694.2 | 1833058H1 | 1959 | 2171 |
| 52 | 435694.2 | g3003709 | 1979 | 2171 |
| 52 | 435694.2 | g3959558 | 1980 | 2162 |
| 52 | 435694.2 | g2358889 | 2057 | 2171 |
| 52 | 435694.2 | g4264824 | 2059 | 2162 |
| 52 | 435694.2 | 474448H1 | 2096 | 2171 |
| 52 | 435694.2 | 4615243H1 | 1 | 250 |
| 52 | 435694.2 | 2533844H1 | 21 | 246 |
| 52 | 435694.2 | g3154852 | 56 | 491 |
| 52 | 435694.2 | 4270549H1 | 86 | 338 |
| 52 | 435694.2 | 3330903H1 | 114 | 331 |
| 52 | 435694.2 | 4315068H1 | 219 | 423 |
| 52 | 435694.2 | 4315053H1 | 221 | 486 |
| 52 | 435694.2 | 4706408H1 | 242 | 486 |
| 52 | 435694.2 | 1436837F1 | 243 | 695 |
| 52 | 435694.2 | 5900458H1 | 243 | 486 |
| 52 | 435694.2 | 5903501H1 | 242 | 486 |
| 52 | 435694.2 | 5903201H1 | 243 | 486 |

TABLE 4

| SEQ ID NO: | Template ID | Component ID | Start | Stop |
|---|---|---|---|---|
| 52 | 435694.2 | 5426328H1 | 1199 | 1361 |
| 52 | 435694.2 | g1989927 | 1223 | 1525 |
| 52 | 435694.2 | g1989319 | 1223 | 1632 |
| 52 | 435694.2 | g572779 | 1276 | 1621 |
| 52 | 435694.2 | 4611420H1 | 1351 | 1610 |
| 52 | 435694.2 | 4243682H1 | 1370 | 1625 |
| 52 | 435694.2 | 5868132H1 | 1375 | 1642 |
| 52 | 435694.2 | 792533H1 | 1412 | 1646 |
| 52 | 435694.2 | g1775555 | 1428 | 1811 |
| 52 | 435694.2 | 1469570H1 | 1472 | 1684 |
| 52 | 435694.2 | 1469570F6 | 1472 | 1787 |
| 52 | 435694.2 | 2461531H1 | 1485 | 1710 |
| 52 | 435694.2 | g2220895 | 1488 | 1893 |
| 52 | 435694.2 | g1196070 | 1498 | 1783 |
| 52 | 435694.2 | 3561975H1 | 1527 | 1843 |
| 52 | 435694.2 | 1830391H1 | 1583 | 1847 |

# Table 5

| Program | Description | Reference | Parameter Threshold |
|---|---|---|---|
| ABI FACTURA | A program that removes vector sequences and masks ambiguous bases in nucleic acid sequences. | PE Biosystems, Foster City, CA. | |
| ABI/PARACEL FDF | A Fast Data Finder useful in comparing and annotating amino acid or nucleic acid sequences. | PE Biosystems, Foster City, CA; Paracel Inc., Pasadena, CA. | Mismatch <50% |
| ABI AutoAssembler | A program that assembles nucleic acid sequences. | PE Biosystems, Foster City, CA. | |
| BLAST | A Basic Local Alignment Search Tool useful in sequence similarity search for amino acid and nucleic acid sequences. BLAST includes five functions: blastp, blastn, blastx, tblastn, and tblastx. | Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410; Altschul, S.F. et al. (1997) Nucleic Acids Res. 25:3389-3402. | *ESTs:* Probability value= 1.0E-8 or less<br>*Full Length sequences*: Probability value= 1.0E-10 or less |
| FASTA | A Pearson and Lipman algorithm that searches for similarity between a query sequence and a group of sequences of the same type. FASTA comprises as least five functions: fasta, tfasta, fastx, tfastx, and ssearch. | Pearson, W.R. and D.J. Lipman (1988) Proc. Natl. Acad Sci. USA 85:2444-2448; Pearson, W.R. (1990) Methods Enzymol. 183:63-98; and Smith, T.F. and M.S. Waterman (1981) Adv. Appl. Math. 2:482-489. | *ESTs:* fasta E value=1.06E-6<br>*Assembled ESTs*: fasta Identity= 95% or greater and<br>Match length=200 bases or greater; fastx E value=1.0E-8 or less<br>*Full Length sequences:*<br>fastx score=100 or greater |
| BLIMPS | A BLocks IMProved Searcher that matches a sequence against those in BLOCKS, PRINTS, DOMO, PRODOM, and PFAM databases to search for gene families, sequence homology, and structural fingerprint regions. | Henikoff, S. and J.G. Henikoff (1991) Nucleic Acids Res. 19:6565-6572; Henikoff, J.G. and S. Henikoff (1996) Methods Enzymol. 266:88-105; and Attwood, T.K. et al. (1997) J. Chem. Inf. Comput. Sci. 37:417-424. | Score=1000 or greater; Ratio of Score/Strength = 0.75 or larger; and, if applicable, Probability value= 1.0E-3 or less |
| HMMER | An algorithm for searching a query sequence against hidden Markov model (HMM)-based databases of protein family consensus sequences, such as PFAM. | Krogh, A. et al. (1994) J. Mol. Biol. 235:1501-1531; Sonnhammer, E.L.L. et al. (1988) Nucleic Acids Res. 26:320-322. | Score=10-50 bits for PFAM hits, depending on individual protein families |

# Table 5 (cont.)

| Program | Description | Reference | Parameter Threshold |
|---|---|---|---|
| ProfileScan | An algorithm that searches for structural and sequence motifs in protein sequences that match sequence patterns defined in Prosite. | Gribskov, M. et al. (1988) CABIOS 4:61-66; Gribskov, M. et al. (1989) Methods Enzymol. 183:146-159; Bairoch, A. et al. (1997) Nucleic Acids Res. 25:217-221. | Normalized quality score≥GCG-specified "HIGH" value for that particular Prosite motif. Generally, score=1.4-2.1. |
| Phred | A base-calling algorithm that examines automated sequencer traces with high sensitivity and probability. | Ewing, B. et al. (1998) Genome Res. 8:175-185; Ewing, B. and P. Green (1998) Genome Res. 8:186-194. | |
| Phrap | A Phils Revised Assembly Program including SWAT and CrossMatch, programs based on efficient implementation of the Smith-Waterman algorithm, useful in searching sequence homology and assembling DNA sequences. | Smith, T.F. and M.S. Waterman (1981) Adv. Appl. Math. 2:482-489; Smith, T.F. and M.S. Waterman (1981) J. Mol. Biol. 147:195-197; and Green, P., University of Washington, Seattle, WA. | Score= 120 or greater; Match length= 56 or greater |
| Consed | A graphical tool for viewing and editing Phrap assemblies. | Gordon, D. et al. (1998) Genome Res. 8:195-202. | |
| SPScan | A weight matrix analysis program that scans protein sequences for the presence of secretory signal peptides. | Nielson, H. et al. (1997) Protein Engineering 10:1-6; Claverie, J.M. and S. Audic (1997) CABIOS 12:431-439. | Score=3.5 or greater |
| Motifs | A program that searches amino acid sequences for patterns that matched those defined in Prosite. | Bairoch, A. et al. (1997) Nucleic Acids Res. 25:217-221; Wisconsin Package Program Manual, version 9, page M51-59, Genetics Computer Group, Madison, WI. | |

## CLAIMS

**What is claimed is:**

1. An isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of:

a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52,

b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52,

c) a polynucleotide sequence complementary to a),

d) a polynucleotide sequence complementary to b), and

e) an RNA equivalent of a) through d).

2. An isolated polynucleotide of claim 1, comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-52.

3. An isolated polynucleotide comprising at least 60 contiguous nucleotides of a polynucleotide of claim 1.

4. A composition for the detection of expression of diagnostic and therapeutic polynucleotides comprising at least one of the polynucleotides of claim 1 and a detectable label.

5. A method for detecting a target polynucleotide in a sample, said target polynucleotide having a sequence of a polynucleotide of claim 1, the method comprising:

a) amplifying said target polynucleotide or fragment thereof using polymerase chain reaction amplification, and

b) detecting the presence or absence of said amplified target polynucleotide or fragment thereof, and, optionally, if present, the amount thereof.

6. A method for detecting a target polynucleotide in a sample, said target polynucleotide comprising a sequence of a polynucleotide of claim 1, the method comprising:

a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide, and

b) detecting the presence or absence of said hybridization complex, and, optionally, if present,

the amount thereof.

7. A method of claim 5, wherein the probe comprises at least 30 contiguous nucleotides.

8. A method of claim 5, wherein the probe comprises at least 60 contiguous nucleotides.

9. A recombinant polynucleotide comprising a promoter sequence operably linked to a polynucleotide of claim 1.

10. A cell transformed with a recombinant polynucleotide of claim 9.

11. A transgenic organism comprising a recombinant polynucleotide of claim 9.

12. A method for producing a diagnostic and therapeutic polypeptide, the method comprising:

a) culturing a cell under conditions suitable for expression of the diagnostic and therapeutic polypeptide, wherein said cell is transformed with a recombinant polynucleotide of claim 9, and

b) recovering the diagnostic and therapeutic polypeptide so expressed.

13. A purified diagnostic and therapeutic polypeptide (DITHP) encoded by at least one of the polynucleotides of claim 2.

14. An isolated antibody which specifically binds to a diagnostic and therapeutic polypeptide of claim 13.

15. A method of identifying a test compound which specifically binds to the diagnostic and therapeutic polypeptide of claim 13, the method comprising the steps of:

a) providing a test compound;

b) combining the diagnostic and therapeutic polypeptide with the test compound for a sufficient time and under suitable conditions for binding; and

c) detecting binding of the diagnostic and therapeutic polypeptide to the test compound, thereby identifying the test compound which specifically binds the diagnostic and therapeutic polypeptide.

16. A microarray wherein at least one element of the microarray is a polynucleotide of claim 3.

17. A method for generating a transcript image of a sample which contains polynucleotides, the method comprising the steps of:

a) labeling the polynucleotides of the sample,

b) contacting the elements of the microarray of claim 16 with the labeled polynucleotides of the sample under conditions suitable for the formation of a hybridization complex, and

c) quantifying the expression of the polynucleotides in the sample.

18. A method for screening a compound for effectiveness in altering expression of a target polynucleotide, wherein said target polynucleotide comprises a polynucleotide sequence of claim 1, the method comprising:

a) exposing a sample comprising the target polynucleotide to a compound, and

b) detecting altered expression of the target polynucleotide.

19. A method of claim 6 for toxicity testing of a compound, further comprising:

c) comparing the presence, absence or amount of said target polynucleotide in a first biological sample and a second biological sample, wherein said first biological sample has been contacted with said compound, and said second sample is a control, whereby a change in presence, absence or amount of said target polynucleotide in said first sample, as compared with said second sample, is indicative of toxic response to said compound.

SEQUENCE LISTING

<110> INCYTE GENOMICS, INC.
HODGSON, David M.
LINCOLN, Stephen E.
RUSSO, Frank D.
SPIRO, Peter A.
BANVILLE, Steve C.
BRATCHER, Shawn R.
DUFOUR, Gerard E.
COHEN, Howard J.
ROSEN, Bruce
CHALUP, Michael S.
HILLMAN, Jennifer L.
JONES, Annisa L.
YU, Jimmy Y.
GREENAWALT, Lila B.
PANZER, Scott R.
ROSEBERRY, Ann M.
WRIGHT, Rachel J.
DANIELS, Susan E.

<120> MOLECULES FOR DIAGNOSTICS AND THERAPEUTICS

<130> PT-1022 PCT

<140> To Be Assigned
<141> Herewith

<150> 60/137,109; 60/137,337; 60/137,258; 60/137,260; 60/137,113;
60/137,161; 60/137,417; 60/137,259; 60/137,396; 60/137,114;
60/137,173; 60/137,411; 60/147,436; 60/147,549; 60/147,377;
60/147,527; 60/147,520; 60/147,536; 60/147,530; 60/147,547;
60/147,824; 60/147,541; 60/147,542; 60/147,500

<151> 1999-06-02; 1999-06-03; 1999-06-02; 1999-06-02; 1999-06-02;
1999-06-01; 1999-06-03; 1999-06-02; 1999-06-03; 1999-06-02;
1999-06-02; 1999-06-03; 1999-08-04; 1999-08-05; 1999-08-04;
1999-08-05; 1999-08-05; 1999-08-05; 1999-08-05; 1999-08-05;
1999-08-05; 1999-08-05; 1999-08-05; 1999-08-05

<160> 52

<170> PERL Program

<210> 1
<211> 756
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 061149.1.j

<220>
<221> unsure
<222> 95
<223> a, t, c, g, or other

<400> 1
gagaaatcca acaagctgct gctagctttg gtgatgctct tcctatttgc cgtgatcgtc 60
ctccaatacg tgtgccccgg cacagaatgc cagcnnctcc gcctgcaggc gttcagctcc 120
ccggtgccgg acccgtaccg ctcggaggat gagagctccg ccaggttcgt gccccgctac 180
aatttcaccc gcggcgacct cctgcgcaag gtagacttcg acatcaaggg cgatgacctg 240
atcgtgttcc tgcacatcca gaagaccggg ggcaccactt tcggccgcca cttggtgcgt 300
aacatccagc tggagcagcc gtgcgagtgc cgcgtgggtc agaagaaatg cacttgccac 360

cggccgggta agcgggaaac ctggctcttc tccaggttct ccacgggctg gagctgcggg 420
ttgcacgccg actggaccga gctcaccagc tgtgtgccct ccgtggtgga cggcaagcgc 480
gacgccaggc tgagaccgtc cagcctgcag cagaatcacc tggcgaggag gtgcctgttg 540
aaaaagataa gaagtttcat tcagaaaaga tgaccctgaa gtaaggcaat gagcttccaa 600
aagtcttctc agtgcaatca ccaggcaaat tttataaggc agtcaagaag tttacctaga 660
aataaaaaaa ctcaagagac aataatagtg attaaagtta aatttgatcc cgttcccata 720
taaatcaatt atttatacac ttaacatcat ttgaga 756

<210> 2
<211> 982
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 404508.3.j

<220>
<221> unsure
<222> 206
<223> a, t, c, g, or other

<400> 2
gagagaggat gactgcgcga gaagaagcta gcttacgaac acttgaaggc agacgacgtg 60
ccaccttgct tagcgcccgt caaggaatga tgtctgcacg aggagacttc ctaaattatg 120
ctctgtctct aatgcggtct cataatgatg agcattctga tgttcttcca gttttggatg 180
tttgctcatt gaagcatgtg gcatangttt ttcaagcact tatatactgg attaaggcaa 240
tgaatcagca gacaacattg gatacacctc aactagaacg caaaaggacg cgagaactct 300
tggaactggg tattgataat gaagattcag aacatgaaaa tgatgatgac accaatcaaa 360
gttagtgcac agaaaatctg ctaatctact tcaagaaatg gctgcctcag tagttcccct 420
tcaagctttt taacttcatt attttctgtc tgggtgggaa tttctctttt cttctctatg 480
tctcatttct tacattttct gtccattttt gcttacacat tcttgcgtta tacatttcac 540
aactaaggat gtattaaaga aatgttataa actttccaac aatatttttt cttctatctt 600
ttcttgtatt tgtcttagta ttatgtttac ctttactctt aatatactct tattacattt 660
aataaaatct tagcaaaacc gttgttaata ttgatgcaaa atagatttaa tcaagaaaat 720
gagaattacc agttacagct cccaaatatc agaacgctag caattaatac aatgaaaaga 780
gaaaatatcc caactccaga cttttctcca aaagttttaa tatcacactt tggggtctag 840
tggagtttat ttgtgggttg cagttatttc aaatgggact aaataatgca ttctccagtt 900
tcttattatc aactagtact tttattcagc ctatggaaag acagttggcc taattgctgt 960
tttgagaacc cagccatagc tc 982

<210> 3
<211> 1039
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 441227.2.j

<400> 3
gcggcgccgg ggattgggag ggcttcttgc aggctgctgg gctggggcta agggctgctc 60
agtttccttc agcggggcac tgggaagcgc catggcactg cagggcatct cggtcatgga 120
gctgtccggc ctggccccgg gcccgttctg tgctatggtc ctggctgact tcggggcgcg 180
tgtggtacgc gtggaccggc ccggctcccg ctacgacgtg agccgcttgg gaccggggca 240
agcgctcgct agtgctggac ctgaagcagc cgcggggagc cgccgtgctg cggcgtctgt 300
gcaagcggtc ggatgtgctg ctggagccct tccgccgcgg tgtcatggag aaactccagc 360
tgggcccaga gattctgcag cgggaaaatc caaggcttat ttatgccagg ctgagtggat 420
ttggccagtc aggaagcttc tgccggttag ctggccacga tatcaactat ttggctttgt 480
caggtgttct ctcaaaaatt ggcagaagtg gtgagaatcc gtatgccccg ctgaatctcc 540
tggctgactt tgctggtggt ggccttatgt gtgcactggg cattataatg gctcttttg 600
accgcacacg cactgacaag ggtcaggtca ttgatgcaaa tatggtggaa ggaacagcat 660
atttaagttc ttttctgtgg aaaactcaga aatcgagtct gtgggaagca cctcgaggac 720
agaacatgtt ggatggtgga gcacctttct atacgactta caggacagca gatggggaat 780
tcatggctgt tggagcaata gaaccccagt tctacgagct gctgatcaaa ggacttggac 840
taaagtctga tgaacttccc aatcagatga gcatggatga ttggccagaa atgaagaaga 900
agtttgcaga tgtatttgca aagaagacga aggcagagtg gtgtcaaatc tttgacggca 960

cagatgcctg tgtgactccg gttctgactt ttgaggaggt tgttcatcat gatcacaaca 1020
aggaacgggg ctcgtttat 1039

<210> 4
<211> 1673
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 277927.2

<400> 4
ggcagttgta aagtcgctgg ccagctagtg gagtggagac tgcagaggga gataaagaga 60
gagggcaaag aggcagcaag agatttgtcc tggggatcca gaaacccatg ataccctact 120
gaacaccgaa tcccctggaa gcccacagag acagagacag caagagaagc agagataaat 180
acactcacgc caggagctcg ctcgctctct ctctctctct ctcactcctc cctccctctc 240
tctctgcctg tcctagtcct ctagtcctca aattcccagt cccctgcacc ccttcctggg 300
acactatgtt gttctccgcc ctcctgctgg aggtgatttg gatcctggct gcagatgggg 360
gtcaacactg gacgtatgaa ggcccacatg gtcaggacca ttggccagcc tcttaccctg 420
agtgtggaaa caatgcccag tcgcccatcg atattcagac agacagtgtg acatttgacc 480
ctgatttgcc tgctctgcag ccccacggat atgaccagcc tggcaccgag cctttggacc 540
tgcacaacaa tggccacaca gtgcaactct ctctgccctc taccctgtat ctgggtggac 600
ttccccgaaa atatgtagct gcccagctcc acctgcactg gggtcagaaa ggatccccat 660
gtgggtcaga acaccagatc aacagtgaag ccacatttgc agagctccac attgtacatt 720
atgactctga ttcctatgac agcttgagtg aggcttgctg agaggcctca gggcctggct 780
gtcctgggca tcctaattga ggtgggtgag actaagaata tagcttatga acacattctg 840
agtcacttgc atgaagtcag gcataaagat cagaagacct cagtgcctcc cttcaaccta 900
agagagctgc tccccaaaca gctggggcag tacttccgct acaatggctc gctcacaact 960
ccccccttgct accagagtgt gctctggaca gtttttttata gaaggtccca gatttcaatg 1020
gaacagctgg aaaagcttca ggggacattg ttctccacag aagaggagcc ctctaagctt 1080
ctggtacaga actaccgagc ccttcagcct ctcaatcagc gcatggtctt tgcttctttc 1140
atccaaggat cctcgtatac cacaggaaga agaggctgga aaaccgaaag agtgtggtct 1200
tcacctcagc acaagccacg actgaggcat aaattccttc tcagatacca tggatgtgga 1260
tgacttccct tcatgcctat caggaagcct ctaaaatggg gtgtaggatc tggccagaaa 1320
cactgtagga gtagtaagca gatgtcctcc ttcccctgga catctcctag agaggaatgg 1380
acccaggctg tcattccagg aagaactgca gagccttcag cctctccaaa catgtaggag 1440
gaaatgagga aatcgctgtg ttgttaatgc agagaacaaa ctctgtttag ttgcagggga 1500
agtttgggat ataccccaaa gtcctctacc ccctcacttt tatggccctt tccctagata 1560
tactgcggga tctctcctta ggataaagag ttgctgttga agttgtatat ttttgatcaa 1620
tatatttgga aattaaagtt tctgacttta aaaaaaaaaa aaaaaaaaaa aaa 1673

<210> 5
<211> 968
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 475311.1

<400> 5
ccgcgtcgat cctgggatgg aggaggtggc ggccgctgag gctcggcgtg aagacggcgg 60
gcatggtggg gcgggagaaa gagctctcta tacactttgt tcccgggagc tgtcggctgg 120
tggaggagga agttaacatc cctaatagga gggttctggt tactggtgcc actgggcttc 180
ttggcagagc tgtacacaaa gaatttcagc agaataattg gcatgcagtt ggctgtggtt 240
tcagaagagc aagaccaaaa tttgaacagg ttaatctgtt ggattctaat gcagttcatc 300
acatcattca tgattttcag ccccatgtta tagtacattg tgcagcagag agaagaccag 360
atgttgtaga aaatcagcca gatgctgcct ctcaacttaa tgtggatgct tctgggaatt 420
tagcaaagga agcagctgct gttggagcat ttcctcatct acattagctc agattatgta 480
tttgatggaa caaatccacc ttacagagag gaagacatac cagctcccct aaatttgtat 540
ggcaaaacaa aattagatgg agaaaaggct gtcctggaga acaatctagg agctgctgtt 600
ttgaggattc ctattctgta tggggaagtt gaaaagctcg aagaaagtgc tgtgactgtt 660
atgtttgata aagtgcagtt cagcaacaag tcagcaaaca tggatcactg gcagcagagg 720
ttccccacac atgtcaaaga tgtggccact gtgtgccggc agctagcaga gaagagaatg 780
ctggtaagaa ggattcctga gtcctgtctt agcgaaggtc cgctttgtct tttccatgct 840
tgaactttca cagctgtact tggagtgtta ctgagtgaaa gccaaaagtg ctttttaaa 900

actaggagac caaacaaaag tagtttacat atacactgta ttcatgaaga ataaaaatat 960
tatgctct 968

<210> 6
<211> 3968
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 013039.2

<400> 6
cccaaagcaa aagaaataca gttacgatgg ggacccccctc tggttgatgg tggatcaccc 60
atttcctgtt acagtgtgga aatgtctcct atagaaaaag atgaacctag agaagtttac 120
caaggttctg aagtagaatg tacagtgagc agccttcttc ctggaaagac atacagcttc 180
agactacgtg cagctaacaa aatggggttt ggaccatttt cagaaaaatg tgatattact 240
acagcccctg ggccaccaga tcagtgcaag ccccctcaag tgacatgtag atctgcaact 300
tgtgcacaag tgaattggga ggttcctttg agtaatggaa cagatgtcac tgaatatcga 360
ctggagtggg gaggagttga aggaagtatg cagatatgtt actgtgggcc tggtctcagt 420
tatgaaataa aaggactttc accagcaact acctattatt gcagggtcca ggctctgagt 480
gttgtgggtg caggcccttt cagtgaagta gtagcctgtg tgactccacc atcagttcct 540
ggcattgtga cctgtcttca agaaataagc gatgatgaga tagaaaatcc ccattattca 600
ccttctacat gccttgcaat aagctgggaa aagccttgtg atcatggttc ggaaatcctt 660
gcctacagca tagactttgg agataaacaa tccctaacag tgggaaaggt tacaagctat 720
attatcaaca atttgcaacc agatacaaca tacagaatac gaattcaagc cttgaatagc 780
cttggagctg gtcctttcag ccatatgata aaattaaaaa ctaagcctct ccctcctgat 840
ccacctcgtc tggaatgtgt tgcctttagc caccagaacc ttaagctgaa atggggagaa 900
ggaactccaa agacattgtc aaccgattct attcagtacc accttcagat ggaggataag 960
aatggacggt ttgtatccct atacagagga ccatgtcata catacaaagt acaaagactt 1020
aatgagtcaa catcctataa attctgtatt caagcttgta atgaagctgg ggaaggtccc 1080
ctctcccaag aatatatttt cactactcca aaatctgtcc cagctgcctt gaaagccccc 1140
aaaatagaga aagtaaatga tcacatttgt gaaattacat gggagtgttt acagccaatg 1200
aaaggtgatc cagttattta cagtcttcaa gttatgttgg gaaaagattc agaattcaaa 1260
cagatttaca agggtcccga ctcttccttc cggtattcca gccttcagct gaactgtgaa 1320
tatcgcttcc gtgtatgtgc cattcgccag tgccaagact ctctgggaca ccaggacctc 1380
gtaggtccct acagcaccac agtgctcttc atctctcaga ggactgaacc accagccagc 1440
accaacagag acactgtgga aagcacaagg acccgacggg cactgagtga cgagcagtgt 1500
gctgccgtca tccttgtgct gtttgctttc ttttccattt tgattgcctt tatcattcag 1560
tactttgtaa tcaagtgaaa atataacttt atttttaac tctattacat tttattttgt 1620
catgtactaa aattatttct gtattgcttt tataaaaaac agtggcattt agcactggca 1680
ttgagactat agcacatcat ttttgccatt ttcagtgctt atattgttag gtagaggctg 1740
gcactttatt agaatgcaag ccacaaaaat atcaattttg ttttttttgt tagggtgggt 1800
cttctttttt tctttccctc tctcttttt taacaaatgc cttcttatag aaaaactttc 1860
taagaggcaa caatttagaa tggatatttt gacgaatcgg catgagtgta acagtgataa 1920
cctgatgctg tttgttttaa agattattac caagtgaaaa attcagaatg aatagaattt 1980
acactaacat gctatataaa atgttaaagt ctgatgctgt gaaagcaatc tagtgctata 2040
tttctacctc ctcatttgtc ttaattattt ggtaagtggg attatgatga gtaactggag 2100
gggcttagaa acaaaaactg gatgaaagag tatgcatgaa gaaaagcttc tttgataaat 2160
gtggagttct tcattataaa tatatattca tgaattcaca gataagtact taaagaacag 2220
acagtttact tggcctaaaa atattttgat gtttactcaa aaagtacctc ttcaggtctt 2280
gagaacatgg aaaagaattg agtgctttta aatacttttt agaaagtaat ccataaaagt 2340
aaattgaatt tcaaacctat ttggcttctg ttttgtgaac ctttgaacta tatgtatgtg 2400
tataagggta tacacatacc atatatggca tataacaagt gtacacatat acacataaca 2460
agtgtagaag tatatattac ataacataca ctcactctgt ctggtatagg ctaattttga 2520
agaactccca taagtttctg ctgcttctcc cattaactgc ttgccaccac catcagaatt 2580
cataatcaaa cctaaccttt ttgtttgggg caccaaatct gaagacaaaa ttaatttgca 2640
ccagtaaact tcaagctgct ttctttcttg aaaactaaac gtttaacgta taatgtctgt 2700
ttggatactg ttccaaattg ttgattgcat gtggttaatg ttgcattaga gcactttgca 2760
attgcataat tcattaatgt tttgtgagct tgcatttgtg agttattgga tgatcagact 2820
gaattttgtc aagtatcaca ttgtacatct tgcctagatg tcgatgactg caagtaataa 2880
tacagtttat aatgaaacta tctacaattc ttgttttagc acatctgtta tccgtaaaac 2940
acctgtaact agcttttta atttattatt tgaattttag gatagcgaat cactaatttt 3000
tagttgctga ggttggcatt ttagtgatta ttaagcactt ctgtcagtct ttgaaaaaag 3060
aacgtatttt ttgtggcttt gaagatctct gaagaatttc ttttataata gaatgggcat 3120
gtattgtaac agttttatgt caaatgatct gtgctgtaga aaaacattaa cccttgttca 3180
aaaaagaaat ggataaactt ggcctttcta agtggtaaga atgacctgtc actataatat 3240

actgtatgtt tacattttat ttaaatttaa tctcttatgt atagggtgat aaccttcccc 3300
agaaacaaca gtgattgcga ttgttttcta gaaacttctt taaagtgcca catttggcag 3360
tacaaatgag tctgagtgta atagcccaga gatttatata tagttgaatg tctaaaatgg 3420
taaaatgtgc cactgtgtca agttacagtg gcttatgttt ttcatagtaa ttcaaatgaa 3480
cttcctattt ttgatagtaa atgtcattta atagtatact tgccatttga gcctcactgc 3540
aaaattagtg cagaggagaa aacaattttt aatgtaatct tgattttacc tcatatactg 3600
tacattccaa aaactctaaa ctttttaaag attatagata cactaccaaa catatcacct 3660
taaaattgta taaggctgaa tgaacttcat acaaatgaaa aaaatctcat aaaaatacat 3720
aaactatgta gcaaaagtat ctgtaaaatc catggaaaat aaaagttgta tcattctttt 3780
tgagatacgt ttattgtatt catatatatt cattatttgc tacctgttta agaaagtgaa 3840
atgttatggt ctcccctctt ccaatgagct taaaacattt ttcccaacag tatataaatc 3900
ttcaacatga gaggatgtat atttattata taaagcccag taaagaataa aattagaagt 3960
tttatcct 3968

<210> 7
<211> 1937
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 238005.4

<220>
<221> unsure
<222> 9, 31, 1913, 1919-1920, 1924, 1927-1928
<223> a, t, c, g, or other

<400> 7
ccgttttcnc aatttgggct ggacgaaaaa nacggtcttg ctttcccggt cgccgctgtc 60
gggaagggct gcagggtgtc cgcgagaccc gcggcccggc gagctgaccc cgcctcgcct 120
ttcctgccta gccctcattc cacggagccg gtcgcgcccg gtccttgcgc gacgcttccc 180
ggcccaggcc gcctggtctg gcgctggagg ccggagtccc gcggcctgtg ctggatccgc 240
gcacacccag tggcggcgga tgggcggccg gggcggccgg ggcggccggt cctgagcgcg 300
gcccgggctg tcagggctgg ctgctggcgg gatggacacc ctggaggagg tgacttgggc 360
caatgggagc acagcgctac ccccaccct ggcaccaaac atcagtgtgc ctcatcgctg 420
cctgctgctg ctctacgaag acattggcac ctccagggtc cggtactggg acctcttgct 480
gctcatcccc aatgtgctct tcctcatctt cctgctctgg aagcttccat ctgctcgggc 540
gaagatccgc atcacctcca gccccatttt tatcaccttc tacatcctgg tgtttgtggt 600
ggcgctggtg ggcattgccc gggccgtggt atccatgacg gtgagcacct cgaacgctgc 660
aactgttgct gataagatcc tgtgggagat cacccgcttc ttcctgctgg ccatcgagct 720
gagtgtgatc atcctggggc ctggcctttg gccacctgga gagtaagtcc agcatcaagc 780
gggtgctggc catcaccaca gtgctgtccc tggcctactc tgtcacccag gggaccctgg 840
agatcctgta ccctgatgcc catctctcag ctgaggactt taatatctat ggccatgggg 900
gccgccagtt ctggctggtc agctcctgct tcttcttcct ggtctactct ctggtggtca 960
tccttcccaa gaccccgctg aaggagcgca tctccctgcc ttctcggagg agcttctacg 1020
tgtatgcggg catcctggca ctgctcaacc tactgcaggg gctggggagt gtgctgctgt 1080
gcttcgacat catcgagggg ctctgctgtg tagatgccac aaccttcctg tacttcagct 1140
tcttcgctcc gctcatctac gtggctttcc tccggggctt cttcggctcg gagcccaaga 1200
tcctcttctc ctacaaatgc caagtggacg agacagagga gccagatgta cacctacccc 1260
agccctacgc tgtggcccgg cgggagggcc tggaggctgc aggggctgct ggggcctcag 1320
ctgccagcta ctcgagcacg cagttcgact ctgccggcgg ggtggcctac ctggatgaca 1380
tcgcttccat gccctgccac actggcagca tcaacagcac agacagcgag cgctggaagg 1440
ccatcaatgc ctgagggcag ctgccagggc ctgtggagga caggccagag aggaggccag 1500
caggccccag agtccccagg ggaggaggac caggtcaagg gacgttctgt gggcagtagc 1560
cctgtgtggc cctgttccca ccatgagtct ggaggcccca cctccctggg gctcccaatc 1620
ccctttgcca tctctgctct cactggggac cctcctcccc ttcccacctg ctctcatact 1680
gctcagtgac atggcccagg ctttccttcc agggccatgc ttggcaaggt tggctgaggg 1740
caccctcctt ctctgcaccc ttggcacgag ggcagggctg gctctcccaa tgcctccatc 1800
ccatccccat ggtgctttgg cctcctcaaa gcatccacca tggtggatgg actgaagtgt 1860
gtatattttc ttgatctatt ttttaataaa aaggaaaagg agcagaaaaa aanaaaaann 1920
aaanaanngg aaaaaaa 1937

<210> 8
<211> 794
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 345322.1.j

<220>
<221> unsure
<222> 50, 779, 786
<223> a, t, c, g, or other

<400> 8
attagcgtat tgttcctttc tgtattgtgc tgagaggatc caagggattn gtgggggaac 60
aggcaagcca ggcatcaccg tggatgttga agaagggggt tactcaaacc tcggcatctt 120
cacttgctcg atctggaatt acagctattt attacgaagc actctgtgtg gcttagtgga 180
gtgtgtctga ggaaacacat cccggacacc acttagggtt agtctttctg agctcttcac 240
atctctgact aattatcatc attaccatgg agcccaacag tcccaaaaag atacagtttg 300
ccgtgcctgt attccagagt cagattgcac ctgaagcagc agagcagatc aggaaaagaa 360
gacctacacc agcatcactt gtgattctca atgagcataa ccccccagaa atagatgaca 420
agagggggcc caacacacaa ggggaattac agaatgcatc ccctaagcaa aggaagcaga 480
gtgtgtacac accacccacc ataaaagggg ttaagcatct gaaaggccag aatgaatcag 540
cattccctga agaagaagaa ggcaccaatg aaagagagga gcagcgggac cattaattac 600
tggtctgcag caagaaggct tcttggaaat aactgaacta ttaacttttc tgagtatacc 660
atggaattcc actgcttgac ttccagaagc atcctccatc tctgcacccc acactcatac 720
agtagctatg cacatcctgg aagtctcctt gactgaactt tagaactaag tacacattnt 780
ccacancact tata 794

<210> 9
<211> 3991
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 348094.6.j

<220>
<221> unsure
<222> 642
<223> a, t, c, g, or other

<400> 9
tttgtagaga ccgtgtctcc ctatgtcgcc caggaaaagg aaactattaa taagcttgga 60
cgaatatgct tcaactctag aattctgaaa aaaatcagaa aactaaaaac ttctgaagct 120
ctatttgaaa tctctcttca aattgagttc taaagttcct gttgcttcag acaatggatg 180
agcaatcaca aggaatgcaa gggccacctg ttcctcagtt ccaaccacag aaggccttac 240
gaccggatat gggctataat acattagcca actttcgaat agaaaagaaa attggtcgcg 300
gacaatttag tgaagtttat agagcagcct gtctcttgga tggagtacca gtagctttaa 360
aaaaagtgca gatatttgat ttaatggatg ccaaagcacg tgctgattgc atcaaagaaa 420
tagatcttct taagcaactc aaccatccaa atgtaataaa atattatgca tcattcattg 480
aagataatga actaaacata gttttggaac tagcagatgc tggcgaccta tccagaatga 540
tcaagcattt taagaagcaa aagaggctaa ttcctgaaag aactgtttgg aagtattttg 600
ttcagctttg cagtgcattg gaacacatgc attctcgaag antcatgcat agagatataa 660
aaccagctaa tgtgttcatt acagccactg gggtggtaaa acttggagat cttgggcttg 720
gccggttttt cagctcaaaa accacagctg cacattcttt agttggtacg ccttattaca 780
tgtctccaga gagaatacat gaaaatggga tacaacttca aatctgacat ctggtctctt 840
gggctgtcta ctatatgaga tggctgcatt acaaagtcct ttctatggtg acaaaatgaa 900
tttatactca ctgtgtaaga agatagaaca gtgtgactac ccacctcttc cttcagatca 960
ctattcagaa gaactccgac agttagttaa tatgtgcatc aacccagatc cagagaagcg 1020
accagacgtc acctatgttt atgacgtagg caaagaggat gcatgcatgc actggcaagc 1080
agctaaaaca ttgcaagatc atgaagagtg taaccaaagt aattgaaagt attttgtgca 1140
agtcgtacct ccccatttat gtctggtgtt aagattaata tttcagagct agtgtgcttt 1200
gaatccttaa ccagttttca tataagcttc attttgtacc agtcacctaa atcacctcct 1260
tgcaaccccc aaatgacttt ggaataactg aattgcatgt taggagagaa aatgaaacat 1320
gatggttttg aatggctaaa ggtttataga atttcttaca gttttctgct gataaattgt 1380
gtttagatag actgtcagtg ccaaatattg aaggtgcagc ttggcacaca tcagaataga 1440
ctcatacctg agaaaaagta tctgaacatg tgacttgttt ctttttttagt aatttatgga 1500
cattgagatg aacacaattg tgaacttttg tgaagatttt atttttaaac gtttgaagta 1560
ctagttttag ttcttagcag agtagttttc aaatatgatt cttatgataa atgtagacac 1620

aaactatttg agaaacattt agaactctta gcttatacat tcaaaatgta actattaaat 1680
gtgaagattt ggggacaaaa tgtgagtcag acactgaaga gttttttgtt ttgttttaat 1740
atttttgata ttctctttgc attgaaatgg tataaatgaa tccatttaaa aagtggttaa 1800
ggatttgttt agctggtgtg ataataattt ttaaagttgc acattgccca aggcttttt 1860
tgtgtgtttt tattgttgtt tgtacatttg aaaaatattc tttgaataac cttgcagtac 1920
tatatttcaa tttctttata aatttaagtg cattttaact cataattgta cactataata 1980
taagcctaag tttttattca taagttttat tgaagttctg atcggtcccc ttcagaaatt 2040
tttttatatt attcttcaag ttactttctt atttatattg tatgtgcatt ttatccatta 2100
atgtttcata ctttctgaga gtataatacc cttttaaaag atatttggta taccaatact 2160
tttcctggat tgaaaacttt ttttaaactt tttaaaattt gggccactct gtatgcatat 2220
gtttggtctt gttaaagagg aagaaaggat gtgtgttata ctgtacctgt gaatgttgat 2280
acagttacaa tttatttgac aaggttgtaa ttctagaata tgcttaataa aatgaaaact 2340
ggccatgact acagccagaa ctgttatgag attaacattt ctattgagaa gcttttgagt 2400
aaagtactgt atttgttcat gaagatgact gagatggtaa cacttcgtgt agcttaagga 2460
aatgggcaga atttcgtaaa tgctgttgtg cagatgtgtt ttccctgaat gctttcgtat 2520
tagtggcgac cagtttctca cagaattgtg aagcctgaag gccaagagga agtcactgtt 2580
aaaggactct gtgccatctt acaaccttgg atgaattatc ctgccaacgt gaaaacctca 2640
tgttcaaaga acacttccct ttagccgatg taactgctgg ttttgttttt catatgtgtt 2700
tttcttacac tcatttgaat gctttcaagc atttgtaaac ttaaaaaatg tataaagggc 2760
aaaaagtctg aacccttgtt ttctgaaatc taatcagtta tgtatggttt ctgaagggta 2820
attttatttt ggaataggta aaggaaacct gttttgtttg tttttcctga gggctagatg 2880
catttttttt ctcacactct taatgacttt taacatttat actgagcatc catagatata 2940
ttcctagaag tatgagaaga attattctta ttgaccatta atgtcatgtt catttaatg 3000
taatataatt gagatgaaat gttctctggt tggaacagat actctctttt tttttcttgc 3060
aatctttaag aatacataga tctaaaattc attagcttga cccctcaaag taacttttaa 3120
gtaaagatta aagcttttct tctcagtgaa tatatctgct agaaggaaat agctgggaag 3180
aatttaatga tcagggaaat tcattatttc tatatgtgga aacttttgc ttcgaatatt 3240
gtatcttttt aaatctaaat gttcatattt ttcctgaaga aaccactgtg taaaaatcaa 3300
attttaattt tgaatggaat aatttcaaag aactatgaag atgatttgaa gctctaattt 3360
atatagtcac ctataaaatg ttctttatat gtgttcataa gtaaatttta tattgattaa 3420
gttaaacttt tgaattgatt tgaggagcag taaaatgaaa gctatatcta ttctaaacct 3480
tatttagaca ttggtaccag ttacccaggt gaaaatatgg agtaactttg ttttgtatgg 3540
taaggtttag gaatggtgga tgaagggtat ctctatataa ataaagtgct caacaatgtg 3600
caatgattgt aaatttagta agatattaca gccatttcat gaatgcttta ccattcaaca 3660
tagtatctat tacaaaacac ctttcttgta tccatatact tcaggtgttg ctgttaacat 3720
ttactatgat atttatttta accaaaatgt tactcacatt aaatgtttat tctttaaaat 3780
gaatgtatta tgtttttaac ccacaaatgc atacttaccc tgtgcctcat atttcaatag 3840
tactgtaata tggacatctt ttgtgaaata cttttatttt gttatgcttt aaatatacat 3900
acaaaaagat ttctgttatt agctttgaaa attgtataat atcctaatat aaacaaaaat 3960
ataaaaataa aaatgaatac agtaaaatgt c 3991

<210> 10
<211> 2885
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 233678.1

<400> 10
aaatagtaag aaaataatgt acagttggta ctcagatgtg gggcctaagc tatgaatggg 60
actagagcat tgctacaggt tgggcattgc tgggcactgt atgtaggtcg cctgacccta 120
tctcccatgg gagtgatgac cagtcccaga ggcctgtgtg actctgaatc ttccaacctc 180
acaggccatt ccctcccttc tgaaagccca cccatgggct tctggcccct gtccaaggcg 240
tgggagggag gccacggcct ctgatgcacc tgggcgtccc ctcccatccc ccctccctgc 300
acaggtgctg gaaacatgca tgaagagctg cggcaagcgg ttccacgacg aagtgggcaa 360
gttccgcttt ctcaacgagc tcatcaaggt cgtgtctccc aagtatctgg gctctcggac 420
atcggagaag gtgaagaaca agatcttgga gctcctctac agctggacag tgggcctgcc 480
cgaggaggtg aaaatcgcag aggcctacca gatgctaaag aagcagggga ttgtaaagtc 540
cgaccccaag cttccagatg acactacctt tcccccttcct cctccacggc cgaagaatgt 600
gatctttgaa gatgaggaga aatccaagat gctggcccgc ctgctgaaga gctcccatcc 660
cgaagacctc cgcgcaccaa taagctcatc aaagagatgg tgcaggagga ccagaagcgg 720
atggagaaga tctcgaagag gggtgaatgc catcgaggag gtgaacaaca atgtgaaact 780
ggctcacgga gatggtgatg agccacagcc agggcggcgc acagctggca gcagcgagga 840
cctcatgaag gaactgtacc agcgctgtga gcggatgcgg cccacgctct tccgactggc 900
gagtgacaca gaggacaatg atgaggcctt agcggagatc ctgcaggcca atgacaacct 960

cacccaggtg atcaacctgt ataagcagct ggtgcggggt gaggaggtca acggtgatgc 1020
cacagccggc tccatccctg gggagcacct cggccctgct ggatctctca ggcctggatc 1080
tcccgcctgc gggcaccacc tacccagcta tgcccacccg ccctggcgag caggccagcc 1140
ctgagcagcc cagtgcctca gtttccctgc ttgacgacga gctcatgtct ctgggcctca 1200
gtgaccccac acccccttca ggcccaagcc tggatggtac cggatggaac agcttccagt 1260
cgtcggatgc cactgagccc ccagcccctg ctctgggccc aggcccccag tatggaaagc 1320
cgaccccccag cgcagacatc cctgccagca agcagcggtc tggacgacct agacctcctg 1380
gggaagaccc tcctgcagca gtcgctgccc ccggaatccc agcaagtgcg gtgggagaag 1440
cagcagccaa ccccccggct cacactccgg gacctgcaga ataagagcag cagctgcagc 1500
tcccccagct ccagcgccac cagccttctc cacaccgtgt ccccagagcc ccccaggcct 1560
ccgcagcagc ccgtaccaac cgagctctca ctggccagca tcactgtgcc cctggagtcc 1620
atcaaaccca gcaacatcct gcccgtgact gtgtatgacc agcacggctt ccgcatcctc 1680
ttccattttg cccgggaccc actgccaggg cgctccgacg tgctggtggt ggtggtttcc 1740
atgctgagca ccgcccccca gcccatccgc aacatcgtgt tccagtcagc tgtccccaag 1800
gttatgaagg tgaagctgca gccaccctcg ggcacggagc tgccagcttt taaaccccat 1860
cgtccacccc tcagcaatca cccaggtcct gctgcttgcc aaccccccaga aggagaaggt 1920
tcgcctccgc tacaagctca ccttcaccat gggtgaccag acctacaacg agatggggga 1980
tgtggaccag ttcccccccac ctgaaacctg gggtagcctc tagaacagag gggctgggga 2040
gaggaagggg cagagggacc ggtcactgtc cagcctggag ggaggcattg gtggccaagg 2100
acacccttttg ttgcccatgg ccattcaccc ccaggcctgg tgcttctccc cacacccctg 2160
taggcctcaa gtgactcttc cccctcctgc tccggccccg cccctgctga gccaaaccca 2220
gtaggaggct ggggcctggg tttgtgccgc tggggtctcc atcaccggga cctggagagg 2280
gaggggctgt gtagccttgg aagaacttgg gtcatgggga ggaagcacag ctgttgggga 2340
agggccagga cctcaggccc agccccaacc ccagctgggg tggggtcttc cccacctgtc 2400
tcttatgcct tatgggaagg cccagccata actcgggggc catgctggag ctggggacca 2460
gcttaggcct cctccatagg aacccagtga ctgggggtg acgcctacac ccccagctat 2520
ttgcactctg gtgtgtggtt tgactctgct tttcttccgg attggccctg tggtcacagc 2580
ctcagggggc cagggctggg ggaacctcac ctggcccgta ctcctggggg tttccctttg 2640
ccattgggcc ccctgaggga ctgtgggggc tcaagggtaa tgccagaggc ccatggcccc 2700
agcgaggggc tgtggggcac ctagagttct cggtgtgtct ccttcattca ttggcctctg 2760
ctggggcctc ctatgggtgt cttacgtctg tccatccatc tgtccgtggt cagaagtggg 2820
gtcagtgtgt gagtgagagc aggagtattt atgaaaataa aacgtcgttt ttcctggaaa 2880
aaaaa 2885

<210> 11
<211> 2458
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 312243.1

<400> 11
ggggcctcac aggccacgtg agcttctgag agctctgagc atgaggtcac gggctcccag 60
aggcagtgat gtggcagcgc ctgtctcccg ccccccgtcc agctccgacg ccggcctgcg 120
gttcccggac agcaacggcc tcctgcagac cccacgctgg gacgagccgc agcgggtgtg 180
cgccctggag cagatttgcg gcgtgttccg cgtggatctg ggccacatgc gctccctccg 240
ccttttcttc agcgacgagg cctgcaccag cggccagctg gtcgttgcca gccgagagag 300
ccagtacaag gttttccact tccaccacgg cggcctggac aagctgtctg acgtgttcca 360
gcagtggaaa tactgcaccg agatgcagct caaagaccag caggtcgccc ccgataagac 420
atgcatgcag ttctccatcc gccgccccaa gctgccgtcc tccgagacgc accccgagga 480
gagcatgtac aagaggctcg gcgtctccgc ctggctcaac cacctgaatg agctgggcca 540
ggtggaggag gagtacaagc tgcggaaggc cattttcttt ggcggtattg atgtgtcaat 600
ccgcggggag gtctggccct tcctgctgcg ctattacagc cacgagtcca cgtcggagga 660
gcgggaggcg ctgcggctgc agaagcgaaa ggagtactct gagatccagc agaaaaggct 720
ctccatgact cccgaggagc acagagcgtt ctggcgtaat gtgcagttca ctgtggacaa 780
agacgtggtc cggacagatc ggaacaacca gttcttccgg ggggaagaca atcccaatgt 840
ggagagcatg aggaggatcc tgctgaacta cgccgtgtac aaccctgccg tcggtattcc 900
caagggatgt cggacctggt ggcgcccatc ttggccgagg tcctggatga gtcagacacc 960
ttctggtgct ttgtgggttt gatgcagaac acgatcttcg tcagctcacc ccgggacgag 1020
gacatggaga aacaactgct gtacctgcgc gagctgctgc ggctgacgca cgtgcgcttc 1080
taccagcacc tggtctcgct gggcgaggac ggcctgcaga tgctcttctg ccaccgctgg 1140
ctccttctgt gcttcaagcg ggagttcccc gaggccgaag cgctgcggat ctgggaggcc 1200
tgctgggccc actaccagac ggactacttc caccttttca tctgcgtggc catcgtggcc 1260
atctacgggg atgacgtcat cgagcagcag ctggccacgg accagatgct cctgcacttc 1320
ggaaacctgg ccatgcacat gaacggggag ctcgttctcc ggaaggcgag gagtttgctg 1380

taccagttcc gcctcctgcc ccggatcccc tgcagcctgc acgatctgtg taagctgtgc 1440
gggtcaggca tgtgggacag cggctccatg cccgcggtgg agtgcaccgg ccaccatccc 1500
ggctcggaga gctgtccctа cgggggcacg gtggagatgc cttcccccaa gtccctgagg 1560
gaaggcaaga agggcccaaa gacgccgcag gacggcttcg gcttccgcag ataggtcggg 1620
cccccgacac cggacagggg ttgaggggac ctcctcagag gccctgggca cgggaggggg 1680
tggggctggg cgtgaagggg acaggggacg atagaaacct aaggaaaatg cttttgggca 1740
acatgagagg aacctttтca tattaatgac aaaattagag tctggaagtg acagaagtca 1800
gatctacagc cacccagagg aaagtcagct cctgaaacgc tgcagtggaa cgcgcagcca 1860
ccgcacctga gacgcaggct ggctgggctc tcctgctggc tgccctggag gatttcaaca 1920
tgtcccagga tttgctccac cctcgagggc agccagacag cgtcgccagg caatgaggaa 1980
agcagagaca ggagaggaag gcctcactca cccactgcgt cgagggctgc agaacacagc 2040
ggggtcctgt ccaggcccag ggacatcttt gcaagccaga cacacttcct cttgagacct 2100
cgttctctcg gagtgagcca aacacacttc ccaaaacgtc cccagccaca gctgggatgc 2160
cgatggaaag gcatctgcca taaaagaaaa gcaaaagata aaaagcccaa ccgatgtggg 2220
gatagagagg cggaagagca gtcaggcttg aggagctggc gcttgtaatg tttatccgtt 2280
taaacatttc gtcctcctgg tacacgaagg gaactgtctg cccaggagcc tgagcctcag 2340
gctgttggag aagcatctga tgccttttтc tttgctgggg gtcttctacg tgaggttcct 2400
tggcgttgtt taaggtcaac tccaccaaat acagcaccca gctggggctt gaatggga 2458

<210> 12
<211> 748
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 425487.3

<220>
<221> unsure
<222> 635
<223> a, t, c, g, or other

<400> 12
gcggccgcag cgtcggggct ggagcgatgg cggcgaccgc ggtggcggcg gctgtggcgg 60
gaaccgagtc ggcccagggt cccccgggcc cggcagcgtc gctggagctg tggctgaaca 120
aagccacaga cccaagcatg tcggaacagg attggtcagc tatccagaat ttctgtgagc 180
aggtgaacac tgaccccaat ggccccacac atgcgccctg gctactggcc cacaagatcc 240
agtctccgca agagaaggaa gctctttatg ccttaacggt gagtttggct tgccttgtag 300
cctaaccttt cctgtccttg tgctatagag agccgagagg cgctttgctt ccacacagat 360
ccttttcact ggagaaggga ggtcccctga gaggtttacc ttcacagcag gttggaggag 420
ggagatctgg gccagggtcc ccacccttct tcctcctgtt tctcgtctgt ttcttcaggg 480
cagcattgag tctactgtgg gagaaggaaa ggaaggtttt atttcatttg cccttcttag 540
accagggctg gcaagttcat gtgccttttc tagagagggc agttctagca cagggcccac 600
tagctgttgg caggtatgag tatgggctca gggtngctag tttaggtctt tgtatatgtt 660
tgagtgttta gatgttggtg actaatttgc agtgttgaaa gccacagtat ggacgagatg 720
aggggtctgc aggctcagct tgcctctt 748

<210> 13
<211> 1098
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 346813.3.j

<220>
<221> unsure
<222> 20, 47, 97, 106, 114, 941, 944, 947, 985, 1014, 1042
<223> a, t, c, g, or other

<400> 13
cacaaaacca ttgaagtttn caacacccta acccagccac cttactnctg gtaacagaga 60
gcccagttaa acataactgt ttagaggtgc tggactnagt ttattntagt aggnctaacc 120
tccgagacca cccttaaaca tcagtagact gggagctgta cgtggatggg agcggctttg 180
ccaaccccctg caaagtgact ctgaagaagg agacaagccc tgctccagtc acacccagaa 240

gctgactggt ccacgcacag gcgaagcatg aggaaactca ttgcgggact cattttcctt 300
aaaatttgga cttgtacagt aaggacttca actgaccttc ctcagattga gaactgtttc 360
cagtatatac atcaagtcac tgaggtagga caaaaattgc tacagtccta ttattttatg 420
gttattataa gtgtaccagg actctaaaag aaacttgttt gtataatgct atccaaggta 480
tgtagcccag ggaataacca acctgatgtg tgttatgacc cattttaagc ctcccgtgat 540
cacagttttt aaaataaaat taaggactgg tccttttcta ggtgacacaa gtaaggtaat 600
agctagaatg gaaaaaagag gggcccccaa aaatgtaacc ttaaaatttg gtgcttgtgc 660
cgctattgat agtaagcagc atggaatagg atgcggttct ctaaattgga aaaaaaaagt 720
gacacagtaa aaaaaaaata agtgtatctg tcaagaattg tatttatgtg agatgtgtca 780
atactggtct tgtgtcattt gggctactta aaaagaagat aaaaaagatc ctgtttggcg 840
gcttagttgt cttgggcaaa aactgtataa tggtaccaca aaaagttaca tggtggagtt 900
ccaattacac agaaagaaat ccattcagta aatttccaaa nttncanact gtttgggccc 960
acccagaact ccaccgggac tgganagccc ccaccgggtt atacgcagag cttntgctaa 1020
gctccctgat cagtggacag gnagctgtgt aattggcacc attaagccat ctttcttctt 1080
atgcccataa aaacaggt 1098

<210> 14
<211> 539
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 006861.1.j

<400> 14
tttgctgttt gatattggaa tacactctta aatcaatgtg gttatgccac acatcatttt 60
aatgtgaatt tctcacttta tgctttttg cgactgagtt attacttgcc gtttaagttt 120
attttagact gtggaaaaga cgttagacaa aaagcaaatt cgagtgattt tcttattcga 180
gttaagaatg ggttgtaaag cagtggagac aactcggcaa catcagcacc acattgggcc 240
caggagccgc tgggaaccac taaacgagca tacagtgcag tgctgggtca agaagtttta 300
caggctggga ttagaggcgt gagcactgtg ccctactgaa cgctgtttct agtgataact 360
tgctgctcta tttttagcaa gtttactata tttatttact ttattcaacc agcaattact 420
gagaatctgc tataggcagg ttctgggaac acaggaaaca aagcaaacaa cagaaataag 480
aattggtctt ttcctggtag gaatttaaaa tctaatggga gcagcagcag tctctttcc 539

<210> 15
<211> 863
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 028008.3.j

<220>
<221> unsure
<222> 522
<223> a, t, c, g, or other

<400> 15
gggatctggc ctttgtaccc ctaggctgga gttaggtccg gagtttttat cttcatcttt 60
agtactgtac tgtttctcag gttcattgag gaggcactga gaacaaatgg gatatttttt 120
ccatgatgag ggaaaaaacc ccaaccctga gagagattca gtgaggatag cgggagacgg 180
aaaagttaga cctcagagca cttctctgcc agctcttcta ttctttttc aagtcccttg 240
tgggcaagga tgtggtcgtg gaactaaaga atgacctgag catctgtgga accctccatt 300
ctgtggatca gtatctcaac atcaaactaa ctgacatcag tgtcacagac cctgagaaat 360
accctcacat gttatcagtg aagaactgct tcattcgggg ctcagtggtc cgatacgtgc 420
agctgccagc agatgaggtc gacacacagt tgctacagga tgcggcaagg aaggaagccc 480
tgcagcagaa acagtgatgg ctcctcctcc tcttcccctc cntctttcat tggtgaccca 540
taaccccaag tcccagccca gaacccctaa cccccaatac ttgaaggggt tttgtttttt 600
tactaatgat ggttttgtgg gttttttta agggatgagt ggatgagagg agtaataggg 660
aacagctatc ctctcttgag aaggggagga taagtaggct gggaaacttc aaagccttcc 720
cagtccccag cacctgcctt tctcactact tctctggaga tggtaggaga gtttcctagg 780
tctttccagg gcagcatgtg attcatttgg ggatggaagg aatctgtccc gcatcgggaa 840
taaaatttat gatgcaaaaa aaa 863

<210> 16
<211> 2600
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 346078.5.j

<400> 16
cctgaggact cagcgaaggg tgggcgccgc cgaggcctcc tgccgctggc gggtttccgc 60
ggagtgccgc ccggctccgc tctgccgccg gcgcggctca tgggcagagt cggccgggcg 120
ggccggcatt aaactgaaga aaagatgtcc ctgtacgatg acctaggagt ggagaccagt 180
gactcaaaaa cagaaggctg gtccaaaaac ttcaaacttc tgcagtctca gcttcaggtg 240
aagaaggcag ctctcactca ggcaaagagc caaaggacga aacaaagtac agtcctcgcc 300
ccagtcattg acctgaagcg aggtggctcc tcagatgacc ggcaaattgt ggacactcca 360
ccgcatgtag cagctgggct gaaggatcct gttcccagtg ggttttctgc aggggaagtt 420
ctgattccct tagctgacga atatgaccct atgtttccta atgattatga gaaagtagtg 480
aagcgccaaa gagaggaacg acagagacag cgggagctgg aaagacaaaa ggaaatagaa 540
gaaagggaaa aaaggcgtaa agacagacat gaagcaagtg ggtttgcaag gagaccagat 600
ccagattctg atgaagatga agattatgag cgagagagga ggaaaagaag tatgggcgga 660
gtggccattg ccccacccac ttctctggta gagaaagaca aagagttacc ccgagatttt 720
ccttatgaag aggactcaag acctcgatca cagtcttcca aagcagccat tcctccccca 780
gtgtacgagg aacaagacag accgagatct ccaaccggac ctagcaactc cttcctcgct 840
aacatggggg gcacggtggc gcacaagatc atgcagaagt acggcttccg ggagggccag 900
ggtctgggga agcatgagca gggcctgagc actgccttgt cagtggagaa gaccagcaag 960
cgtggcggca agatcatcgt gggcgacgcc acagagaaag gtgtgtcccc agggaagcgt 1020
gtgactagag ggaaaggact ggccccatcc atatcagaca tggccagtct tgatcctcat 1080
gtgtcagcag ggggacaatg aggcgtgtgg ccagagggag agggctggcc ctgccatcac 1140
tagaacacag gccgtcctgt tcatatgatg cactgccact tccgttttgt gaaaccagga 1200
atcctgaggc tcatctttat tttttcagaa cagacgtaga gagatgaagg cttgtggagg 1260
aaaagatggt gagagacttg ggcagaaaat gagtagtcct caggaagaaa tcttggttat 1320
gtgtttagag catgaaggac agagccatat agtgtggcag tgaatatacc tgctatctcc 1380
atctcagagg tcgtctctac ttttcccttt tgcctttca gtatagatgt gatttctgat 1440
tctcttacag attgtttgct ttgcgagatc tgatgttatg ttgcagtctc ttggtaaatg 1500
atgcctagtt ggtgttttat tttcatttaa tttttacagt ctgttctgtg ttgagggaat 1560
tcaggaaaga gacaaacata tgttagcatt ttaatcaggg aattaagttt gagtcagcct 1620
agctgaactt cctttgctaa agaaagaaga aaacttttct ggcagccccg ttcatgcaca 1680
gcttaggata catcacgagc ctgacaggtg agtgccagaa accaacagtt gtcccgactt 1740
gtgaggttat ctgaagtaag gcagccggtg gctggattag taactgcata ttcccctggg 1800
cccgtgacct tgaacgtttg ctccaagtca actcacctat aggaattatc actcacatgc 1860
cctgtcagcc ctttgggaag tgagatgagc aaaaattgca agtaatggtg gaggctcaaa 1920
acatccagat gctattgtaa aaacatgcca aagcaaagca gaggctttta ttgcagataa 1980
ggctgtgttt tcgctcagag accaattgtg tagatgccta ggacataaat ggcggggatc 2040
gctattgaaa ttaaattaat tattgtaagt aggactcagt tctgtaacac atctaatgat 2100
atgctgctca gttctgtaac acatctaatg gtatgttttg atacagatgc atccaagaag 2160
tcagattcaa atccgctgac tgaaatactt aagtgtccta ctaaagtggt cttactaagg 2220
aacatggttg gtgcgggaga ggtggatgaa gacttggaag ttgaaaccaa ggaagaatgt 2280
gaaaaatatg gcaaagttgg aaaatgtgtg atatttgaaa ttcctggtgc ccctgatgat 2340
gaagcagtac ggatattttt agaatttgag agagttgaat cagcaattaa aggttagtgg 2400
tacagctaaa tattaaagaa taaaaaagtt gaatttacag ccttaatctt tacaagtaaa 2460
gttactgttt aattaaagtt aaacctttat cttataggat acctgtatta actgtctttt 2520
gtttgccttt cagcggttgt tgacttgaat gggaggtatt ttggtggacg ggtggtaaaa 2580
gcatgtttct acaatttgga 2600

<210> 17
<211> 414
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 394637.1.j

<400> 17
tggttatgtt atacataatt ttaacttgca tttcttgctt tatgtttttt ttgctaatga 60
cttactactt gctgtttatt ttatatttat tttagactat ggaaatgatg ttagacaaaa 120

agtaaattcg agagattttc ttatttaagc tcacaatggg tcgtaaaaca gcagagacaa 180
cttgcaatat caacatgcat ttggcccagg aactgctaat gaacgtacag tgcagtggtg 240
gttcaagaag ttttgcaaaa gataccagag ccttgaggat gaggagtgta gtggcaggcc 300
actggaagtt gacaacaacc aattgagagc aatcattgaa gcagattgat cctcttacaa 360
ttacatgaga agttgctgaa gaactcaacg taagaactca acgtcaacca ttcc 414

<210> 18
<211> 1307
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 222429.3

<400> 18
gggaaatcca gccctcccaa actcttgcct ttcttgtaag actttctaag cgcatggctt 60
tgtcgtaaga cttctcgtgg ccatgtgtct caccctaccg ttcgtctcac caatgaacca 120
gagctgggat accagcaaaa aggacccaaa agactgtttc accacaattg aagaggtcgt 180
ggaccatccc ccaaaatgaa tcccaacact ggaagaccaa gagggttgac gtttgctatc 240
ttgaaagctg cagcccccgt agagaagatc ctagagcaga aggagcacag gctgggctgc 300
catagcatgg acccccaaaa agccactgtc atgaagaaat atcctgtgaa gaaaatcttt 360
gcagggcacc tgaactccca aagccactga ggagaagaat tgcaaccttg gccaatggac 420
ctgatcggtt ttggttatgc agccctcgtg acatttggaa gcatttttgg atataagcgg 480
agaggtggtg ttccgtcttt gattgctggt ctttttgttg gatgtttggc cggctatgga 540
gcttaccgtg tctccaatga caaacgagat gtaaaagtgt cactgtttac agctttcttc 600
ctggctacca taatgggtgt gagatttaag aggtccaaga aaataatgcc tgctggtttg 660
gttgcaggtt taagcctcat gatgatcctg agacttgtct tgttgctgct ctgagcatct 720
ggaggaacag aaaactaagt tcatgtcatc ctgctgtaat gggcagagca tatttttttt 780
gtatttaaaa gataaacttc aatatggaat gctagaaaca caaatagcac tgtcacctct 840
aatatgaaca ttagtttgag gtagtttttt tctaaaacaa aaattttaac tgttttctaa 900
ttgtcaagca ctattttcat taaaagtgtc taatgaatca tgatatactc ttccatttgt 960
tgtgtctatt ttttatatat ttggtatttt ttgaaaattc caaatactca tgtctcaagt 1020
aagcttaaac tacaacttgt cacataaagg aagtcttaag tggagttcac agaatgataa 1080
tgtatctatt tgtcatttgt gttatatttg aaattattag aaattatgct ttttccattt 1140
taattgtatt gctgccagtg ctatttttt ctttaaaaaa ttttattctt agcacactgt 1200
tatgtcctaa ctgaatgtat tcagtattca aataaaagac aaatgacaaa tagatacatt 1260
atcattctgt gaactccact taagacttcc tttatgtgac gtcgacg 1307

<210> 19
<211> 1406
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 366739.2

<220>
<221> unsure
<222> 311
<223> a, t, c, g, or other

<400> 19
tgctgccagc gagagccgcg ggagagtgtg cagccgagtc actactgcct gcctgcctgc 60
ctgctacggt gagtgtggcc cccacaatgg gatggcgcag ggcaggaggg ccatgggttc 120
ccccacccca gactaagggg gcactagggg aggggccgag tcatgtgaag agggagaccc 180
tctcagacag tcgaatgtgc tggtcccact aaggaaacca cctcaccctc tccaacttcc 240
tgcctgaaaa tgggccctgg agctcgcaga cagggcagga ttgtgcaggg aaggcctgag 300
atgtgcttct ncccaccccc taccccactc cctccccttc ggatcttaac actgggcact 360
cacacaccca ccccatgctc ctctccaggc tcagcagcag gtactgtacc caaccatggg 420
ctcgcaggcc ctgcccccgg ggcccatgca gaccctcatc tttttcgaca tggaggccac 480
tggcttgccc ttctcccagc ccaaggtcac ggagctgtgc ctgcctgcct gtccacagat 540
gtgccctgga gagccccccc acctctcagg ggccacctcc cacagttcct ccaccaccgc 600
gtgtggtaga caagctctcc ctgtgtgtgg ctccggggaa ggcctgcagc cctggcagcc 660
agcgagatca caggtctgag cacagctgtg ctggcagcgc atgggcgtca atgttttgat 720
gacaacctgg ccaacctgct cctagccttc ctgcggcgcc agcaccagcc ctggtgcctg 780

gtggcacaca atggtgaccg ctacgacttc cccctgctcc aagcagagct ggctatgctg 840
ggcctcacca gtgctctgga tggtgccttc tgtgtggata gcatcactgc gctgaaggcc 900
ctggagcgag caagcagccc ctcagaacac ggcccaagga agagctacag cctaggcagc 960
atctacactc gcctgtatgg gcagtcccct ccagactcgc acacggctga gggtgatgtc 1020
ctggccctgc tcagcatctg tcagtggaga ccacaggccc tgctgcggtg ggtggatgct 1080
cacgccaggc ctttcggcac catcaggccc atgtatgggg tcacagcctc tgctaggacc 1140
aagccaagac catctgctgt cacaaccact gcacacctgg ccacaaccag gaacactagt 1200
cccagccttg gagagagcag gggtaccaag gatcttcctc cagtgaagga ccctggagcc 1260
ctatccaggg aggggctgct ggccccactg ggtctgctgg ccatcctgac cttggcagta 1320
gccacactgt atggactatc cctggccaca cctggggagt aggccaagaa ggaaaatctg 1380
acgaataaag acccccgctg ccccat 1406

<210> 20
<211> 3028
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 474635.6

<220>
<221> unsure
<222> 2063-2064, 2068, 2070, 2078-2079
<223> a, t, c, g, or other

<400> 20
ccccagcccg ccatggcgtc aggcgccgcg gccccgggga ggtggctccc actttaagaa 60
gtgaagtttt gctcccctcc ccctccctgc ccacctcctg cagcctcctg cgccccgccg 120
agctggcgga tggagctgcg cacgggagcg tgggcagcca ggcggtggcg cggaggatgg 180
atggggacag ccgagatggc ggcggcggca aggacgccac cgggtcggag gactacgaga 240
acctgccgac tagcgcctcc gtgtccaccc acatgacagc aggagcgatg gccgggatcc 300
tggagcactc ggtcatgtac ccggtggact cggtgaagac acgaatgcag agtttgagtc 360
cagatcccaa agcccagtac acaagtatct acggagccct caagaaaatc atgcggaccg 420
aaggcttctg gaggcccttg cgaggcgtca acgtcatgat catgggtgca gggccagccc 480
atgccatgta ttttgcctgc tatgaaaaca tgaaaaggac tttaaatgac gttttccacc 540
accaaggaaa cagccaccta gccaacggta ttttgaaagc gtttgtctgg agttagaaag 600
ttctcttctt caacacgtcc ctccccaggg tgttcctccc tgtgacccag ccgcctcgac 660
ttcggcccgc ttgctcacga ataaagaact cagagttgtg tgtgcaatgc acacccagac 720
acacgcacgc acacacacgc gcgcgcacac acatgctttt ttctgttccc ctccgctttc 780
tgaagcctgg ggagaaatca gtgacagagg tgttttggtt ttattgttat gtgggttttc 840
ttttgtattt tttttgtttg ttttgttttt aaacattcaa aagcaattaa tgatcagaca 900
taggagaaac cctgaataga aacaaaactt ttgaatgctg gattcaaaaa aaaaaaaaag 960
ttatctggac agcttctttg agactattta aaaactggta caacaggtct ctacaacgcc 1020
aagatctaac taagctttaa aaggtcaaga agttttatgg ctgacaaagg actcgcgcaa 1080
cgcagaaggc ctttcccacc ttaagcttcc ggggatctgg gaattttacc cccattctct 1140
tctgtttgtc tgagtctcat ctctctgcaa gcaagggctg aaatcatttt gtttggttgt 1200
tttgagggag agaggcgggg tgggggggtg caaatctgcc agcagctctt acgtaaggca 1260
tgttttattg gggagggctg agcttttatt ttctcctctc cagtggggtt ggcttttatt 1320
gtttcttgtt tgggtttgga atggaaatat ggatagcagc ataaagtact tttattttga 1380
caaaattcat ttttttcaac aatggagaca tagatttgac ccacaataac ttctcccccc 1440
ctctttttac tctgctcaaa aagcatctct cctcccatta cccaaccttg gtcataagtg 1500
tgcctggctg gtttgcagat atttgttctg ctttgtaaaa attggccatt agtgcattta 1560
ttgagatgat ctctaaagag ctatgccctg acctacccct gattctatga cattggggcc 1620
cttcttttgc tgaaactgcc ttacgtaatg gttttactcc ttgaaagaga tttgacggaa 1680
tccatttttat gccaagtgct gccctgcact gtttctgcaa tatgtggtgt atgctgtggt 1740
gatcttgctg ggaatgatta taagtgtgtg tgtggtgggg gagtgggtat tacatgcatt 1800
gctgaagagt catcctggtg ttcctcattc ctcccacctt cccgtggtca ttttaattac 1860
ggggcagtgt caccgcaaag ggaggaaact caaagccgaa agcaaaattc caggcctgat 1920
tctggctttt gaggttcctg gttcttgaag ccaggcctga cccgactctc agatggggtc 1980
agtcccgtcg ctttgcagac tgaccctgga aatctacaaa atgcagattt tcctgatttc 2040
ctcttctctt gcccagtttt ttnntgtntn tttttttnna aaagcctgga ttgtaaccag 2100
attttctttt ttccccccttc tcagctgtag atatgatatc tcctttcagg gccccagctt 2160
aagggcaaag tgagttaatg tgtagacaaa ggcgagggac aagagagagt taacatctag 2220
acagtggaaa aagccatggt gtgtggtttc tgggaaccac caacacttgc aggtttagct 2280
ttttcccagg gttgactaca agaaagaaaa ccatgttttt gcaagattaa aatgtggttg 2340
agtgtgccta aattaaccat ccccatttttt atcatatttc caccatcact tcagggtttt 2400

aagagtcagt gctcacctgg gcggagctgg tagtacattt tgcttcttag aaagctaagt 2460
cctgggttcc gtctgatttt aggttccagg aacttcctga gaacacccga tcgcagaggg 2520
taattttctg gagtttgttt tgcagggata gctgggagta tggccaccct gctccacgat 2580
gcggtaatga atccagcaga agtggtgaag cagcgcttgc agatgtacaa ctcgcagcac 2640
cggtcagcaa tcagctgcat ccggacggtg tggaggaccg aggggttggg ggccttctac 2700
cggagctaca ccacgcagct gaccatgaac atccccttcc agtccatcca cttcattttt 2760
ttttgcaggg tgctgcctat gggccctctg ctccccaatg ccttagagag aggaggggag 2820
ccacggccgc tcaccggaag gctgtgtgcg gggacatccg aggtggtggt ggacaggaag 2880
gacttgggaa ggggagcgag aaattgcttt ttctcttcct ccttgggcag aatgtagctt 2940
ttctgcttca ctgtggcagc ctcctccctg gatccttaga tcccagagga gggaagaaaa 3000
tttgcagtga ctgaaaacag taaaaaaa 3028

<210> 21
<211> 537
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 228470.1.j

<220>
<221> unsure
<222> 110, 114, 159, 163
<223> a, t, c, g, or other

<400> 21
cggagaagtc atactctctc acaccctcgg ctttcttgtt gtgtccttca gcaaaacagt 60
ggatttaaat ctccttgcac aagcttgaga gcaacacaat ctatcaggan aganagaaag 120
aaaaaaaccg aacctgacaa aaaagaagaa aaagaagang aanaaaaatc atgaaaacca 180
tccagccaaa aatgcacaat tctatctctt gggcaatctt cacggggctg gctgctctgt 240
gtctcttcca aggagtgccc gtgcgcaggg aagatgccac cttccccaaa gctatggaca 300
acgtgacggt ccggcagggg gagagcgcca ccctcaggtg cactattgac aaccgggtca 360
cccgggtggc ctggctaaac cgcagcacca tcctctatgc tgggaatgac aagtggtgcc 420
tggatcctcg cgtggtcctt ctgagcaaca cccaaacgca ttacagcatc gagatccaga 480
acgtggatgt gtatgacgag ggcccttaca cctgctcggt gcagacagac aaccacc 537

<210> 22
<211> 3080
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 407090.5.j

<220>
<221> unsure
<222> 2926, 2938, 2941, 3008, 3052
<223> a, t, c, g, or other

<400> 22
gcggcgctgg ctttaggtga acgacgtggt gaggagtggg tttcgggcat gagaagtcac 60
agggccgttt cctagtctct cttcacttct ttgggtcttc tcagagaaag aaggctgccg 120
tgggtaggct gggggcggag actatcggga agagaaaatt acttttccca ctgaaacaca 180
cccaagtata tgcccagcct tcatgaaagt gaacagagaa acgaagcgcc tttatgtggg 240
tggccttagc caggacattt ctgaggcaga cctacaaaat cagttcagca gatttggaga 300
agtttcggat gtggagatca tcacacggaa agatgaccaa ggaaacccac agaaagtttt 360
tgcatatatc aacatcagtg tagcagaagc ggacctgaaa aaatgtatgt ctgttttaaa 420
taaaacaaaa tggaaaggtg gaacattaca aattcaacta gcaaaagaaa gctttctgca 480
cagattggcc caagagagag aagcagcaaa agctaagaaa gaagaatcaa caacaggtaa 540
cgccaacttg ttagaaaaga caggaggagt ggatttccat atgaaagctg tgccagggac 600
agaagtgcca gggcataaga attgggttgt gagcaaattt ggaagagtct tacctgttct 660
tcaccttaaa aatcaacata aacgtaaaat catcaaatat gatccctcaa aatactgcca 720
caacctgaag aagatagggg aggatttctc aaacaccatt cctatatcca gcctgacttg 780
ggaattagaa ggagggaatg accctatgag taagaaacgg cgaggagagt tctctgactt 840
tcatggccct cccaagaaga taataaaagt gcagaaggat gagagttccc actgggtctc 900

tggccatgag tacaaggccc aggagggtaa tagagagacc acccttaaca cagcaacagg 960
ctgcacaaaa aagaacttgt gattccatta ctccttctaa atcatctcct gtacctgttt 1020
ctgatactca gaaacttaaa aatctacctt ttaagacttc tggcttggaa actgccaaga 1080
agagaaacag catttctgat gatgatactg attctgaaga tgaattgaga atgatgattg 1140
cgaaagagga aaacttacag agaactacac aaccctcaat aaatgaatct gaaagtgatc 1200
cttttgaagt tgtaagggat gatttcaaat caggcgttca caaactgcat tctttaatag 1260
gtttaggtat caaaaatcgt gtctcttgcc atgatagtga tgatgatatt atgagaaatg 1320
atcgtgagta tgactcagga gatacagatg aaattattgc gatgaaaaaa aatgttgcta 1380
aggtcaaaaa cagtacagaa ttttcacaaa tggaaaaatc tacgaagaaa acttctttca 1440
aaaatagaga aaactgtgag ctttctgatc actgtattaa actacaaaaa agaaaaagca 1500
atgtagagtc agccctcagt catggattaa agtctcttaa tcgtaaatct ccctctcact 1560
ccagtaggca gtgaagatgc tgattctgca tcagaattag ctgactctga aggaggtgag 1620
gagtataatg ccatgatgaa aaactgcctt cgtgtgaatc tcactttagc tgatttggaa 1680
caattggctg gcagtgatct gaaggttcca aatgaagata ctaagagtga tggaccagaa 1740
accaccaccc aatgcaagtt tgacagatgc tccaagagcc ccaagactcc cactggcctc 1800
cgcagaggcc gacagtgtat tcgtcctgcg gagattgtgg cttccctgtt agaaggagag 1860
gagaacacct gtggcaaaca gaaaccaaag gaaaacaatt taaagccaaa atttcaggct 1920
ttcaagggag taggctgtct atatgaaaag gagtcaatga aaaaatcctt gaaagacagt 1980
gttgcctcta acaataaaga tcagaattcc atgaaacatg aggatcccag tatcatatcc 2040
atggaagatg ggtccccata tgttaatggc tcattaggtg aagtgactcc atgccaacat 2100
gcaaagaagg cgaatggccc aaactatatt cagcctcaaa aaagacagac cacttttgaa 2160
agccaggatc gcaaggcagt gtcccctagc agttctgaaa agagaagtaa gaatcctatt 2220
tctaggccat tagaaggtaa gaagtcctta agtcttagtg caaagactca caacataggc 2280
tttgacaaag acagctgcca tagtaccaca aagacagaag cttcacagga agagcggtct 2340
gattcaagcg gcctcacatc tctcaagaaa tcaccaaagg tctcatccaa ggacactcgg 2400
gaaatcaaaa ctgatttctc actttctatt agtaattcgt cagatgtgag tgctaaagat 2460
aagcatgctg aagacaatga gaagcgtttg gcagccttgg aagcgaggca aaaagcaaaa 2520
gaagtgcaga agaagctggt gcataatgct ctggcaaatt tggatggtca tccagaggat 2580
aagccaacgc acatcatctt cggttctgac agtgaatgtg aaacagagga gacatcgact 2640
caggagcaga gccatccagg agaggaatgg gtgaaagagt ctatgggtaa aacatcaggg 2700
aagctgtttg atagcagtga tgatgacgaa tctgattctg aagatgacag taataggttc 2760
aaaattaaac ctcagtttga gggcagagct ggacagaagc tcatggattt acagtcgcac 2820
tttggcaccg atgacagatt ccgcatggac tctcgatttc tagaaactga cagtgaagag 2880
gaacaggaag aggtaaatga aaagaaaact gctgaggaag aagagnttgc tgaagaanaa 2940
nagaaagccc tgaatgttgt acaaagtgtt ttgcaaatca acttaagcaa ttctacaaac 3000
agaggatnag tagctgctaa gaaatttaag gacatcatac attatgatcc ancgaagcaa 3060
gaccatgcca cttacgaaag 3080

<210> 23
<211> 426
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 068194.1.j

<220>
<221> unsure
<222> 328, 403
<223> a, t, c, g, or other

<400> 23
cccaagtccc tagaagaagc caccccatcc aaggagggtg acatcctaaa gcctgaagaa 60
gaaacaatgg agttcccgga gggggacaag gtgaaagtga tcctgagcaa ggaggacttt 120
gaggcatcac tgaaggaggc cggggagagg ctggtggctg tggacttctc ggccacgtgg 180
tgtgggccct gcaggaccat cagaccattc ttccatgccc tgtctgtgaa gcatgaggat 240
gtggtgttcc tggaggtgga cgctgacaac tgtgaggagg tggtgagaga gtgtgccatc 300
atgtgtgtcc caacctttca gttttatnaa aaagaagaaa aggtggatga actttgcggc 360
gcccttaagg aaaaacttga agcagtcatt gcagaattaa agntaaacat gtattctgaa 420
aacaaa 426

<210> 24
<211> 3219
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 411449.2

<220>
<221> unsure
<222> 3088
<223> a, t, c, g, or other

<400> 24
ggaaagggaa acaaccacat aaacgcaaga aaaaatccag gaaaaagtct ctcaaaaaac 60
ctgctttatt cttagaggca gaaagtaaca cttcacattc agatgattca gcatccagca 120
gttctgagga aagtgaggaa agagacacta agaaaaccaa aaggaaaaag agagagaaaa 180
aagcccatac ctctgtagcc aacaatgaaa tacaggagag gacaaacaaa cgcacaaatt 240
ggaaagtagc tacagatgaa aggtctgctg agagctcaga ggatgactaa atgggaaaca 300
cttttgtttt ccacatgact gtggatattt acagttctta ctccttgtgg ttttgccagt 360
gactcttgtt cagcacgggg cctgaggtca gagctgtctt gtgccatctg tatgttctga 420
cagacgtctt gtcttctatt ttggcgttaa gcttgatccc cttttcttgt taaaagggaa 480
tctggtattt tgttatgaag gtttcttgaa gagatttttt tttttgcaat taattacgtt 540
tagtgtagag tgcatataca gcaaattaaa ggacccagaa agctggatcc aatagtgacc 600
tgggtacacc aatcggaata ttgaatttgg ggaagtcaag ggctgggatc caagaggtgg 660
attggaacta atgccatgta ggatggtatg acaaggcaac actgtattgc tctctgttta 720
tatagcaggt gtcacaacta acttgtcttt agccttggtg ctttgatcct tctatatttt 780
gaccccacag gtgtggtccg gtttacttaa tcaggacatg ggcctaagaa caaacctttt 840
cccttcatga taacatccat agacaactta ttagaaggga ctagagtttt tgcaaatttc 900
cctgctggat ggggcctata gctatactta gtatatgcct aaacatggta attggatagt 960
aaatggtttt ctagttccat tgctgtatat ttgcctaaat ggacttgtgt tcaaattatt 1020
tcttcaattg tcatagataa tcctgtacca aatggggaag aattaggaaa taatcatgtt 1080
gtctaatggt actctggatt cagggcagca actgccattt aaatgttgtc ttgttcattt 1140
ctaaatctgt tcatgaagtt taggttttcc ctgaaactaa gttgaattat ttccaaaatg 1200
aaacaggctt ctcagggaca tatccacttc ttcccagtct gcctttggat taaagcacca 1260
agcagagacc acattaattc cctttgctat actgtgatcc ttagtatgtt aattcttaag 1320
aaaccaacat atcactgaaa gaaggctggc agaacgcaag tgcatttttt cactgtggga 1380
agaaagatca agtgacgtat tatttttcc tggttgtcac ttaatgggct gagtaaaaag 1440
cttgaaaact cagactttcg gtcttggttc tgccactcat tggttatgag gaggcccaga 1500
gcaggtaagt tcaccttcct ggccttactt tcctgatgtg taatacggaa ttacttcaca 1560
gtagcatgac agtataagac accagcagta gatacaacta tgatgacatt ccatgagttg 1620
gtatttttag ttctaactgc taaatttgtt ctctttacgg gacagatttc taataaagtg 1680
cttggtctta aaatacatgg ttggacagag gtgccctatc ccttaactat gagcaggtgc 1740
tacctttttgg gatatttatt ttaaatttta atactttggt actcaattgt cagtgttcca 1800
tggtgtgtat ttttatttttt gggattagtg ggggtctaaa gggagaagaa tagtctctaa 1860
ttactacctc ttaacctaaa gcaattattt tgttcctgga gcaagttaaa tctttgttgg 1920
aaggagcttt ggccatatat tttttagcat gcattgtttc tgtgccctga aagtacctga 1980
aaggttttaa gcacagactc aggaaaatgt gccagtagaa caggccatct ccaggaaatt 2040
ggctctattt gggtcctgac cttcccttcc tcccaagtta gcaggcttgt tcttttgcaa 2100
ggaatacaca tcttgccttt ttttttttt ttgccatgtt ttccttttct tggtcatgta 2160
taagcaataa agctgttttt tgttcttcat ctttcttaac cccaaatttt cttctatgcc 2220
ttaggcttcg atggttcttc caacccccct aatatggctt agggtggttt ttcaaaacct 2280
acaatccccc atttgcacta ctggccatgg aacatttatt tctagtgttc ctgccaatca 2340
gagatctcta tattaaattc taaaatggga ttaaaagaag agttggagaa ttcacactta 2400
ttgagtaact gatgtcatac aacctggaat ttctgaattc caaataaata aatttcactc 2460
tttgaacatt tcatctttta ctttttagca ccaacagact tgataacagc ctgatgctga 2520
tctgacaatg ggttgatagc cttcccccac tgacccttaa atctgcttag taacaagtcc 2580
tttgcttctg tcatttctcct gggggatggc ctactgccct cctttctgta caatctgggc 2640
aaaccgactg gtgatggcaa gagtggtgtc aatgaagcgg tctacacagc tggagagaca 2700
attttcagtg cgagagtcta ggcgattccc tggcttctcc acacatttat cccaacataa 2760
ctccatgaag tgatgcacta tttgtgaggc tgggtactaa gaaatcagtt ctctctaagg 2820
tcctcaagga tagctgtcat cacctcccat ttaagaggcg tgattatgta gtccaaggtc 2880
atgtagccag caagaagtca ggccgcgtta gaaccatgtc cgaagggctc caaacccttg 2940
ttctacatcc atagtctaca gcgactactt cagagtccac cttcctccga taatgttcta 3000
gtcgttttca aatacattgt cgcatatgat cctcacaatc cagtgaggca gtggggtgac 3060
ggcgaaacca aaacccacag tggaatgnag tatcttttct acgggcacgt ggcctgtgca 3120
gtaaactgcg ccttctgctg ctcggcggcc accaggcgct gcaactccgc ttcatcggct 3180
tcgcccagct ccgccattgt tcgcctgcag gctcgccac 3219

<210> 25
<211> 445

<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 018549.2

<220>
<221> unsure
<222> 34, 418
<223> a, t, c, g, or other

<400> 25
tgacttttgg gagagctgac cttttgtgac tttngggaga gctgccaaaa gtgaaactta 60
gtgcctcaga caagcagggg caagtctgct aaggaagctg tggccagaag cacagatcag 120
aaacacgatg gctctgttaa cagccgaaac attccgctta cagtttaaca acaagcgcct 180
cctcagaagg ccttactacc cgaggaaggc cctcttgtgt taccagctga cgccgcagaa 240
tggctccacg cccaccagag gctactttga aaacaagaaa aagtgccatg cagaaatttg 300
ctttattaac gagatcaagt ccatgggact ggacgaaacg catgcttacc aagtcacctg 360
ttacctcacg tggagcccct gctcctcctg tgcctgggag ctggttgact tcatcaangc 420
tcacgaccat ctgaacctgg gcatc 445

<210> 26
<211> 1657
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 236043.3

<220>
<221> unsure
<222> 5
<223> a, t, c, g, or other

<400> 26
ccggnccggg ctcctcctgc tgctgggaca gtgctctagt agaacagaca gacctactga 60
cacaggggag gtgagaaggg aggtgaccac caggactggc tctgtgagta ccacacagtg 120
gggaggggggt gggggccacc atgtcatcat atcagaagga actggagaaa tacagagaca 180
tagatgaaga tgagatccta aggaccttga gccccgagga gctagagcag ctggactgcg 240
aactacagga gatggatcct gagaacatgc tcctgccagc tggactaaga caacgtgacc 300
agacaaagaa gagcccaacg gggccactgg accgagaggc ccttttgcag tacttggagc 360
aacaggcact agaagtcaaa gagcgtgatg acttggtgcc cttcacaggc gagaagaagg 420
ggaaacccta tattcagccc aagagggaaa tcccagcaga ggagcagatc accctggagc 480
ctgagctgga ggaggcactg gcacatgcca cagatgctga aatgtgtgac attgcagcaa 540
ttctggacat gtacacactg atgagtaaca agcaatacta tgatgccctc tgcagtggag 600
aaatctgcaa cactgaaggc attagcagtg tggtacagcc tgacaagtat aagccagtgc 660
cggatgaacc cccaaatccc acaaacattg aggagatact aaagagggtc cgaagcaatg 720
acaaggagct ggaggaggtg aacttgaata atatacagga catcccaata cccatgctaa 780
gtgagctgtg tgaggcaatg aaggcaaata cctatgtgcg gagcttcagt ctggtagcca 840
cgaggagtgg tgaccccatt gccaatgcag tggctgacat gttgcgtgag aatcgtagcc 900
tccagagcct aaacatcgaa tccaacttca ttagcagcac aggactcatg gctgtgctga 960
aggcagttcg ggaaaatgcc acactcactg agctccgtgt agacaatcag cgccagtggc 1020
ctggtgatgc agtggagatg gagatggcca ccgtgctaga gcagtgtccc tctattgtcc 1080
gctttggcta ccactttaca cagcaggggc cacgagctcg ggcagcccag gccatgaccc 1140
gaaacaatga actacgtgag taactgcaga catgatgtgt ggagtggtca gggaagtgca 1200
gaaattggtg gatcctcctg aaagcagacc taatgactaa cagcccaggg tgctaccaaa 1260
gagctattca tatgagaatt tcaaatccca agagtgatca gaaagtaaac agaaaactcc 1320
ccctgcccca aatatggacc agtagagagg tagaaaggat gccagagata atcatacttg 1380
tttagaggta tcgtaatttt atttgtggtg tggtttcttg ttttgttttg ttaatgtggg 1440
atgggcctgt aaggtggcct aagacactac caatttatga gtttggctaa gggactggaa 1500
aggagaagag tgcctttgca gaaagcagga gctggaacaa acacttatgt ttaatattgc 1560
tcctttacag gtcgccagca aaagaagaga taacactgca tttcccttta ccaactagcg 1620
ctgggagcac tggtcactta aatcctcatc tgtcctc 1657

<210> 27

<211> 1041
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 445433.2

<400> 27
ttcttcttta attagaagct taaagagaag tttgcagaat gtactcataa gtggatggga 60
taatactgtt aagttctgat attctgatat tgtttgaaat actgttaaga atttcacatt 120
tggtaagtat tttttatatc agtattaaaa tagtaatttg gtttattaca attttataca 180
tagaatttgc cagttacttt ctgactacaa agaaaaacag atactaaaag tctcttctga 240
aaacagcaat ccagaacaag acttaaagct gacatcagag gaagagtcac aaaggcttaa 300
aggaagtgaa aatagccagc cagaggaaat gtctcaagaa ccagaaataa ataagggtgg 360
tgatagaaag gttgaagaag aaatgaagaa gcacggaagt actcatatgg gattcccaga 420
aaacctgcct aacggtgcca ctgctgacaa tggtgatgat ggattaattc caccaaggaa 480
aagcagaaca cctgaaagcc agcaatttcc tgacactgag aatgaacagt atcacaggga 540
cttttctggc catcccaact ttcccacgac ccttcccatc aaacagtgat gaacaaaatg 600
atactcagaa gcaactttct gaagaacaga acactggaat attacaagat gagattctga 660
ttcatgaaga aaagcagata gaagtggctg aaaatgaatt ctgagctttc tcttagttat 720
aagaaagaaa aagacctctt gcatgaaaat agtacgttgc aggaagaaat tgtcatgcta 780
agactggaac tagacataat gaaacatcag agccagctaa gagaaaagaa atatttggag 840
gaaattgaaa gtgtggaaaa aaagaatgat gatcttttaa agggtctaca actgaatgag 900
ctcaccatgg gatgatgata ctgccgtgct cgtcattgac aacggctctg gcacgtgcaa 960
ggccggcttt gcaggtgacg atgccccccg ggctgtcttc ccttccatcg tgggggtgcc 1020
cccaggcacc agagcatgat g 1041

<210> 28
<211> 2113
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 344630.7.j

<400> 28
gcagtgcggc cgtcatggcg tcgcccttca gcggggcgct gcagctgacg gacctggatg 60
acttcatcgg gccgtctcag gagtgcatca agcctgtcaa agtggaaaaa agggcgggaa 120
gtggcgttgg ccaagattcg cattgaagat gacgggagct acttccaaat taaccaagac 180
ggcgggaccc ggaggctgga gaaggccaag gtctcgctaa acgactgcct ggcgtgcagc 240
ggctgcatca cctccgcaga gaccgtgctt atcacccagc agagccacga ggagctgaag 300
aaggttctag atgctaacaa gatggcggca cccagtcagc agaggctggt tgtagtttcg 360
gttctcacca cagtctagag catcgctggc tgcacggttt cagctgaatc ctacagatac 420
tgccaggaaa ttaacctcat tctttaaaaa aataggggtg cacttcgtct tcgacaccgc 480
cttctcaagg cacttcagcc tcctgggaga gccagcgaga gtttgtgcgg cgattccgag 540
gacaggccga ctgcagacag gcgctgcccc tgctggcctc tgcctgccca ggctggatct 600
gctatgccga gaagactcac ggcagcttca tcctccccca catcagcacc gcccggtccc 660
cgcagcaggt catgggctcc ctggtcaagg acttcttcgc ccagcagcag cacttgaccc 720
ctgacaagat ctaccacgtc acagtgatgc cctgctatga caaaaagctg gaagcctcca 780
gacccgactt tttcaaccag gagcaccaga cacgggatgt ggactgtgtc ctcacaacag 840
gagaagtttt caggttgctg gaggaagagg gcgtctccct ccccgacctg gaaccagccc 900
ctctggacag cctgtgcagc ggtgcctctg cagaggagcc caccagccat cggggagggg 960
gctcgggggg ctacctggag cacgtgttcc ggcacgcggc ccgagagctc tttggaatcc 1020
atgtggctga ggttacctac aaaccctga ggaacaaaga cttccaggag gtgacactgg 1080
agaaggaggg ccaggtgctg ctgcacttcg caatggcgta cggcttccgc aacatccaga 1140
acctggtgca gaggctcaaa cgagggcgct gcccctacca ctacgtggag gtcatggcct 1200
gcccctcagg ctgcctgaac ggcgggggcc agctccaggc cccagacagg cccagcagag 1260
agctcctcca gcacgtggag agactgtacg gcatggtccg ggctgaggcg cccgaggacg 1320
cgcctggggt tcaggagctg tacacacact ggctgcaggg cacggactcg gagtgtgcag 1380
gtcgcttgct gcatacgcag taccacgccg tggagaaggc cagcactggc ctggtgcatc 1440
cggtggtagg ggctgcagga ccaggactcc caggaggccg tgtccatgtg tgacagcaga 1500
accacatgcc ccaagacccc agggcttccc ccaaaattct gagtgagctg cagggtgtgc 1560
tgggacccga gtaggagcta ggactagcca ggacccgcag ccgcctcgtc acctccagtt 1620
gggtgcctct gggttcccac tggctctgcc caggtggggt ggggtggccc aggcagcaga 1680
aggttccctg aggtcccaga gcctgttccg ttggccctgg gccgaggccc acaggtgctg 1740

cccttgctgc tgctggtcgg gcacccaagt gcgtgagggg cttcagcctg tcccggggtt 1800
gcctgaggca gagcaagacg ggttctcacc cctgacttct ggaggcttcc cttgaagctc 1860
tgtgcaaaag gtgggagaca gagctggacc tgcaggggtg gtcccgccac aaccctgcgt 1920
gtggaccctg gcaggggggg gtgccaggcc cctggaaagc aggggttacc gttacgaggc 1980
tgtggtccgg ggcaagccaa gtacgaagca gcagccatcg cgggctgcat catcccccag 2040
ccaggtcccc accaggcctg tctcccagcg tttgtctaat aaacgcaccc ctcctaaaac 2100
acgcctacga aaa 2113

<210> 29
<211> 3813
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 257121.2

<220>
<221> unsure
<222> 1365
<223> a, t, c, g, or other

<400> 29
aaatcaaggg agctaaatgg cggggtggat ggaattgcaa gtattgaaag tatacattct 60
gaaatgtgta ctgataagaa ctccattttc tctacaaata cctcttctga caatggatta 120
acttccatca gcaaacaaat tggagacttc atagagtgcc ctttgtgcct tttgcggcat 180
tctaaagaca gatttcctga tataatgact tgtcatcaca gatcttgtgt ggattgctta 240
ccgacaatat ttaaggatag aaatctctga aagcagagtt aatattagtt gcccagaatg 300
tactgaacgg tttaatcccc atgatattcg cttgatatta agtgatgatg tcttgatgga 360
aaaatacgaa gaatttatgc ttagacggtg gcttgttgca gatcctgatt gtaggtggtg 420
tccagctcca gactgtggat atgctgtgat agcatttgga tgtgccagct gtccaaaatt 480
aacttgtggg cgagagggct gtggaacaga gttttgctac cactgtaaac agatttggca 540
ccccaaccag acctgtgatg ctgctcgaca agagagagcc cagagcttac gtttgagaac 600
tatacgttct tcatccatta gttatagtca agagtctgga gcagcagctg atgatataaa 660
gccatgtcca cgatgtgctg cttatataat aaagatgaat gatgggagct gcaatcacat 720
gacatgtgct gtttgtggtt gtgagttttg ttggttgtgt atgaaagaaa tctcagattt 780
gcattatcta agtccatcag gatgtacttt ttgggggaag aaaccctgga gccgaaagaa 840
gaaaatattg tggcaactgg gaacactggt tggtgctcct gtcggaatcg ctttaatagc 900
tggcattgct attcctgcaa tgattattgg cattcctgtg tatgtgggcc gcaagattca 960
caatcgctat gaaggcaagg atgtttcaaa gcacaaacgg aatttggcca tagcaggtgg 1020
tgtaacgttg tctgtaatcg tgtctccagt agtagctgca gtgactgtag gtatcggtgt 1080
tcctattatg ttagcttatg tctatggcgt agttccaatt tctctttgtc gaagcggagg 1140
ttgtggagtc tcagcaggca atggaaaagg agttaggatt gaatttgatg atgaaaatga 1200
tataaatgtt ggtggaacta acacagctgt agacacaaca tcagtagcag aagcaagaca 1260
caacccaagc ataggggagg gaagtgttgg tgggctgact ggcagtttga gtgcaagtgg 1320
aagccacatg gatcgaatag gagccatccg agacaacctg agtgnaacgg ccagcaccat 1380
ggcactagct ggagccagta taacggggag tctgtcagga agtgccatgg taaactgttt 1440
taacaggttg gaagtacaag cagatgtaca gaaagaacgg tacagtctaa gtggagaatc 1500
tggcacagtc agcttgggaa cagttagtga taatgccagc accaaagcaa tggcaggatc 1560
cattctgaat tcctacatcc cattggacaa agaaggcaac agtatggagg tgcaagtaga 1620
tattgagtca aagccatcca aattcaggca caacagtgga agcagtagtg tggatgatgg 1680
cagtgccacc cgaagtcatg ctggcggttc atccagtggc ttgcctgaag gtaaatctag 1740
tgccaccaag tggtccaaag aagcaacagc agggaaaaaa tcaaaaagtg gtaaactgag 1800
gaaaaagggt aacatgaaga taaatgagac gagagaggac atggatgcac agttgttaga 1860
acaacaaagc acgaactcaa gtgaatttga ggctccatcc ctcagtgaca gtatgccttc 1920
tgtagcagat tctcactcta gtcatttttc tgaatttagt tgttctgacc tagaaagcat 1980
gaaaacttct tgtagtcatg gttccagtga ttatcacacc cgctttgcta ctgttaacat 2040
tcttcctgag gtagaaaatg accgtctgga aaattcccca catcagtgta gcatttctgt 2100
ggttacccaa actgcttcct gttcagaagt ttcacagttg aatcatattg ctgaagaaca 2160
tggtaacaat ggaataaaac ctaatgttga tttatatttt ggcgatgcac taaaagaaac 2220
aaataacaac cactcacatc agacaatgga attaaaagtt gcaattcaga ctgaaattta 2280
ggcccataaa tgctgcagaa taattaccac tgtacaaccg tgtttggagc tggttgaact 2340
acatgtgact acttaagttt caggttacca gcaaaagccg ggtttcatta tcataatgca 2400
gatacatttt ctgtgttcag caaggcattg tgtgtcatgt gggatcttag ttaccaaact 2460
atgaagtgaa ggctttaaaa gtgcattatt ttaaggataa taaatttgaa gagcaaagca 2520
tgttttgtgt gtttgccaca aaacattgct tgaagcacat acttagatag aaattggtct 2580
taatttatat aatcaatata aaatactaat gcaattctac agcattcaaa tgaagaaaac 2640

ttgaggcttt agggataagt ggttagtgat attttattga aaccactaaa gagataagtt 2700
taaaagaact gcataggtta ctctcagtat atgatactct gtaacatttc tatttatatc 2760
ggcataaatt tcatttttt tcttcatatg caatgtggtt atataaagct taatgcagct 2820
catttgctac catttggata cttagacact ttgagcaaga ttgtggcagt ttttgcacaa 2880
ctttgaaata gaaatacctg gtactctatc ttgtttattg ttgatgccat cttagaggaa 2940
aaaatgtaaa ggtaagtaat taagcatatg acagcaacaa ataagataca taaaactaca 3000
aaataaagtc ccattaggtt ataagtatta caaaaaatcc acctttctct aaggggaagt 3060
ttgtacccca ttgattcttg gtgcctttgg gatcgactgg gttttaatgg cctagttatt 3120
tgaggatttt gctgtgttgt tttccatgtc ttctctggtc accttggatt atatataaaa 3180
atacaggaaa tagataaaca tgaatgtgat taataatgct gaaaaagtat tagcctacca 3240
aagacacact caggctttag tgaataactt tacataacct cagtttttaa cacatgcata 3300
tcttctccaa ccatgaaatc aaagcacggt gcagaacttg taccaagtac aaaaggtcca 3360
tgtatgatta gcattatttt cttttgcttt tgtttatgga caatgttcag ctgacataag 3420
cagaagttgg ccaaaatact gcctgtactg ttaatttcct gtataattca cttaaataaa 3480
agcaggttaa cctcaatgat agcagttaaa atgttctatc ttatgtattt cttttaagta 3540
ttaccattat ggtgctactg agcgttttct tttggtaaaa agaaaaatgc catgggctgc 3600
agtcttcttc catcactttt ccctaccagg tccattaata tgcttataac actagtgcca 3660
gttattttat ttgataatgc ttatggtatt tgtatatttg tttaagtgaa ttatacagga 3720
aattaacagt acaggcagta ttttggccaa cttctgctta tgtcagctga acattgtcca 3780
taaacaaaag caaaagaaaa taatgcgtcg acg 3813

<210> 30
<211> 882
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 243794.1.j

<400> 30
ggcccaaact gagttgcgtt gcctcttctt gcaaatgaac ttgctccgta aaattgagaa 60
cctggaacct ctgcagaaac tggatgctct taacctcagc aacaattaca tcaagaccat 120
tgaaaacctc tcctgcctcc cagtcctgaa cacattgcag atggcccaca atcacctgga 180
gaccgtggag gacattcagc atctacaaga gtgtttgagg ctttgtgtcc ttgacctttc 240
gcacaacaag ctgagtgatc cggaaatatg gcctcaatat gtgccgccag tgtttccgtc 300
agtacgcgaa ggatatcggt ttcattaaga aagacctgag ctgtcttcct tggcactgcc 360
tatggaggtg acacccatct cctccatcat ggccatcctg agaccgctcg cgaagcccaa 420
gatcatcaaa aagagcacca agttcactgg gaaccagtca gactgatatg tcaaaattaa 480
gggtaactgg tggaaacaca gaggtattga caacagggtt catagaaggt ttgagggcca 540
gatctatgcc caacattggt tatgggagaa acaaaaagac aaagcacata ctgcccagtg 600
gcttctggaa gttcctggtc cacaacgtta aggagctgga agtactgctg gtgagcaaca 660
aatcttactg tgttgagatc actcatgatg tttcttccaa gaactgcaaa gccatcttgg 720
aaagagcagc ccaggtggtc atcagagtca ccaatgccaa tgccagcctg cacagtgcag 780
aaagtgaata gacagtgaat gtgtttgttt tattggggtt taaataaaac caataaaact 840
gtaaaagcag cggcaacaaa aaccagaaaa aaaaagtcga cg 882

<210> 31
<211> 514
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 442085.1.j

<400> 31
ctcgttcctg tcgcgcagca cggacctcca cttccacatc tcccccggcg tcggcgcggt 60
cagttgaacc atggcggact ccaaggccac ctcggcggtc accctccgca cccgcaagtt 120
catgaccaac cgcctcctgg cccgcaagca attcgtgctt gaggtgatcc accccggccg 180
cgccaacgtc tccaaggcgg agttgaagga gaggcttgcc aaggcgtacg aggtgaagga 240
ccccaacacc atctttgtct tcaagttccg cacccacttc ggaggaggaa agtccactgg 300
tttcggcctc atctacgaca acctcgaggc tgccaagaag ttcgagccga aataccgcct 360
catcaggaat ggtcttgcta ctaaggttga gaagtcccgc aagcaaatga aggagcggaa 420
gaacagggcc aagaagatcc gtggtgtcaa gaagaccaaa gctggtgacg ccaagaagaa 480
gtaaacgttc gtttacattt gtattactgt tctg 514

<210> 32
<211> 766
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 370661.3.j

<220>
<221> unsure
<222> 48, 354
<223> a, t, c, g, or other

<400> 32
cgggaatccg ccgtttgcgc tgaggcaatg gcggcagctg cgccggtngc cgcggacgac 60
gatgagcggc ggcggcggcc gggggctgca ctggaggact cccggtccca ggaaggggca 120
aatggtgagg ccgagtcagg tgagctcagc cggcttcggg ctgagctggg caggcgccct 180
ggcagaaatg gaaaccatga aggctgtggc agaggtgagc gagagcacga aggccgaggc 240
tgtggctgcg gtgcacggca gtgccaagag gaggtggcct cgctgcaggc catcctgaaa 300
gactccatca gcagctatga agcccagatc accgccctga agcaggagcg acanagcagc 360
agcaggactg tgaggagaag gagcgggagc tgggccgcct gaagcagctg ctgtcccggg 420
cctaccccct ggactcccctg gagaagcaga tggaaaaggt tgggattggg atggcttccc 480
tgctgccctg ggatttggaa gaatgttgag gactgtggat gtcccattcc ctggtcatgg 540
ggtgcggggc acccagcagt gattctagga atagcacaga gttaccaggt gttgttgaga 600
gcttggctct gtgctccgta gggtcctggc ttgcccctta ttagtgggtg accaagttcg 660
cttagctcca aaattaactt aagttatgta gagggaaggg aacttacact taagggtaac 720
aaaaaaaatt cttactgtta ttaaatattg acttctggtg tacatt 766

<210> 33
<211> 1416
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 427939.17.j

<220>
<221> unsure
<222> 1284, 1412
<223> a, t, c, g, or other

<400> 33
atttttttcta aatgtgagat caaggaatta ccaccaaaaa aggagagtaa tacaggagaa 60
atattccaga cagtaatgtt ggaaagacat gaaagccacg acatacaaga tttttgcttc 120
agagaaaccc agaaaaatgt acatgactct cagtgtctgt ggaaacatga ttgaagacat 180
tataagcgag tgcgtgtgac ctataaggaa agtctcattg gtagaagaga catgcatggt 240
agaaaggatg atgcacaaaa gcagcctgtt aaaaatcagc ttggattaaa cccgcagtca 300
catctaccag aactgcagct atttcaagct gaagggaaaa tatataaata tgatcacatg 360
gaaaaatctg tcaacagtag ttccttagtt tccccacccc aacgtatttc ttctactgtc 420
aaaacccaca tttctcatac atatgaatgt aattttgtgg attcattatt cacacaaaaa 480
gagaaagcaa atattgggac agaacactac aaatgtaatg agcgtggcaa ggcctttcat 540
caaggcttac attttactat acatcaaata atccatacta aagagacgca atttaaatgt 600
gatatatgtg gcaagatctt caataaaaaa tcaaaccttg caagtcatca aagaattcat 660
actggagaga agccatataa atgtaatgaa tgtggcaagg tcttccataa tatgtcacac 720
cttgcacagc atcgcaggat tcatactgga gagaaaccat ataaatgtaa tgaatgtggc 780
aaggtcttta atcaaatttc acaccttgca caactatcaa aagaattcat accggagaga 840
aaccttataa atgtaatgaa tgtggaaagg tcttccatca aatttcacac cttgcacaac 900
atcggacaat tcatactgga gaaaaacctt acgaatgtaa caaatgtggc aaggtgttca 960
gtcgcaattc ctaccttgta caacatctga tcattcatac tggagagaaa ccttacagat 1020
gtaatgtatg tggaaaggtc ttccatcata tttcacacct tgcacaacat cagagaatcc 1080
acactggaga gaaaccttac aaatgtaatg agtgtggcaa ggtcttcagt cacaagtcat 1140
ccctagtaaa tcactggaga attcatactg gagagaaacc ttacaaatgt aatgagtgtg 1200
gcaaggtctt cagtcacaag tcatccctag taaatcactg gagaatccac actggagaga 1260
aaccttacaa atgtaatgaa tgtngcaagg tcttcagtcg caattcatac cttgcccaac 1320
atctgataat tcatggccgg tgagaaacct tataagtgtg atgaatgtga caaagcattc 1380

agtcaaaatt cacatgcttg tacaacatca cngaat 1416

<210> 34
<211> 441
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 430569.2.j

<400> 34
aacacttgcg ttgtaatcaa ctacgtgaat aagtgttgtt attcttacct acagaagcca 60
ggttcctgaa gatttcttgc atcacatctc tgtagagatt cttttgagaa ggatccagca 120
aagcccactc ctcctgagta aaggccacag ccacatcttc aaatgacact gaatcctaga 180
atatcgcaca tatgtggaga ggaggatggc ttagacagac agtactaaga atctatactc 240
atctcataaa atcgtaacat aattctgcag acttcaaaca tttcttccat gacctggtca 300
tcagaacttt actctctgtc tgcacttact gctgcccact caacattctt catgctaaaa 360
tgaaactatt cagaaagtca acagcagagt aggaacacct gtctcattgg caggtgcagg 420
gaataaactg tgctgaaaga a 441

<210> 35
<211> 275
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 444689.1.j

<220>
<221> unsure
<222> 44
<223> a, t, c, g, or other

<400> 35
ctatgcaggt tgttgtcggg ataccagatt tcctactccg agangccccg ggtccctctg 60
ccacaacttc tgtcgctctg ccgcctgcac cgtgacccgc actattcacg ggagccctag 120
agaggacacc gggacaccca gaagccggga aatgatgttt caggattcag tggcctttga 180
ggatgtggct gtcagcttca cccaggagga gtgggctttg ctggatcctt cccagaagaa 240
tctctacagg gatgtgatgc aggaaacctt caaga 275

<210> 36
<211> 517
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 445198.1.j

<220>
<221> unsure
<222> 48-49
<223> a, t, c, g, or other

<400> 36
cttgcagaag gtggttatcc tttacggttc tccacactct ttcaggannc agcagaaaat 60
gaacatatct caggcatcag tgtcattcaa ggacgtgact atagaattca cccaggagga 120
gtggcagcaa atggcccctg ttcagaagaa tctgtacaga gatgtgatgc tggagaacta 180
cagcaacctc gtctcagtgg ggtactgctg tttcaaacca gaggtgatct tcaagttgga 240
gcaaggagag gagccttggt tctcagagga ggaattctca aaccagagtc acccaaaaga 300
ttacagaggt gatgacctga tcaagcagaa caagaaaatc aaagacaaac acttggagca 360
agcaatatgt atcaataata aaacattgac tacagaggaa gagaaagttt tggggaaacc 420
atttactctg catgtagctg ctgttgcttc aacaaaaatg tcctgcaaat gcaactcatg 480
ggaagtgaat ttgcaaagta tttctgaatt tatcatt 517

<210> 37
<211> 499
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 084399.1

<220>
<221> unsure
<222> 28, 136, 141, 157
<223> a, t, c, g, or other

<400> 37
ggacatgtcc aggcggaagc aggcgaangc cgctcggtga aagttgaaga gggggaggcc 60
tcagacttct cgctggcctg ggattcctcc gtgacagcag caggaggcct agaaggagag 120
ccagagtgcg atcagnaaac nagccgtgcg ctggaanaca ggaacagcgt gacaagtcaa 180
gaggagagaa atgagggatg atgaagacat ggaggatgaa tcaatttaca cctgcgatca 240
ctgtcagcag gacttcgagt ctctggcaga cctgacggac caccgggccc accgctgtcc 300
tggaggtaat gcaaaacagc cccaggactc tgacaggtgg ttacacgggt aattggacat 360
agcaacagct tgaagcctca agtgagatta aagaattaat aatgtaattt tacagttaat 420
gtcattttat tgtggtgtct tgggtaccct gtatccgcct tttaaataca attaaaaata 480
aaagcaacag ctttccggc 499

<210> 38
<211> 1017
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 350044.1

<220>
<221> unsure
<222> 775
<223> a, t, c, g, or other

<400> 38
ctggcggagg ccttgctgat gaacctgact gagggtcccc tggcgatggc agaaatggac 60
cctacacagg gccgtgtggt ctttgaggac gtggccatat atttctccca ggaggagtgg 120
gggcaccttg atgaggctca gagattgctg taccgtgatg tgatgctgga gaatttggcc 180
cttttgtcct cactaggttc ttggcatgga gctgaggatg aggaggcacc ttcacagcaa 240
ggtttttctg taggagtgtc agaggttaca acttcaaagc cctgtctgtc cagccagaag 300
gtccacccta gtgagacatg tggcccaccc ttgaaagaca ttctgtgcct ggttgagcac 360
aatggaattc atcctgagca acacatatat atttgtgagg cagagctttt tcagcaccca 420
aagcagcaaa ttggagaaaa tctttccaga ggggatgatt ggataccttc atttgggaag 480
aaccacagag ttcacatggc agaggagatc ttcacatgca tggagggctg gaaggactta 540
ccagccacct catgccttct ccagcaccag ggccctcaaa gcgagtggaa gccatacagg 600
gacacagagg acagagaagc ctttcagact ggacaaaatg attacaaatg tagtgaatgt 660
gggaaaacct tcacctgcag ctattcattt gttgagcacc agaaaatcca cacaggagaa 720
aggtcttatg aatgtaacaa atgtgggaaa ttctttaagt acagtgccaa tttcntgaaa 780
catcagacag ttcacactag tgaaaggact tatgagtgca gagaatgtgg aaaatccttt 840
atgtacaact accgactcat gagacataag cgagttcaca ctggagaaag accttatgag 900
tgcagcgaat gccagaaggc ctttattaga aagtctcacc tggttcatca ccagaaaatc 960
cacagtgaag agaggcttgt gtgctccatg aatgtgggaa ttctttagct aaaactc 1017

<210> 39
<211> 1231
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 441329.2

<400> 39
ttttgcttga taataactat aattactaca aggactgcct gagatgaaaa agacaggtga 60
aatcgaactt tcacatggac tggtgtctga gctggagagt ctgtttgagt ttgagtaggg 120
gaaagcctga aacccccccac cacctgctcc cgttagtccc tgcattaact aggggcggac 180
tgggacctga tcctccaagc tggaggcttg aaaactcaag ggcaggcgcc tggtccagcg 240
attcggaaac cgagatgcaa agtcctcccg cgctcctaga gaggacccgg aagcgcggcg 300
cggtcccgga agacgaggtg gtgacacgct tccggccttt gtgactcatt gtgtctgtgt 360
cgaggcgtcg ggagggccta agtccgtgtg cggtgccctt cggccggcct gagccccaga 420
gtcagctccc ctttctcgcc cagcgccccc aggccgctcc cggggctcac ggaatagtaa 480
agaaacacat cataaaacct cccaggacat aaaggtgagc acagaccctg tttggatcaa 540
gtcagttcct ggagcctgaa tgatgactgc tgaatcacgg gaagccacgg gtctgtcccc 600
acaggctgca caggagaagg atggtatcgt tatagtgaag gtggaagagg aagatgagga 660
agaccacatg tgggggcagg attccaccct acaggacacg cctcctccag acccagagat 720
attccgccaa cgcttcaggc gcttctgtta ccagaacact tttggggccc cgagaggctc 780
tcagtcggct gaaggaactt tgtcatcagt ggctgcggcc agaaataaac accaaggaac 840
agatcctgga gcttctggtg ctagagcagt ttctttccat cctgcccaag gagctccagg 900
tctggctgca ggaataccgc cccgatagtg gagaggaggc cgtgaccctt ctagaagact 960
tggagcttga tttatcagga caacaggtcc caggtcaagt tcatggacct gagatgctcg 1020
caaggggat ggtgcctctg gatccagttc aggagtcctc gagctttgac cttcatcacg 1080
aggccaccca gtcccacttc aaacattcgt ctcggaaacc ccgcctctta cagtcacgag 1140
gtaagaagca aggtttcatt taggggaagg gaaatgattc aggacgagag tctttgtgct 1200
gctgagtgcc tgtgatgaag aagcatgtta g 1231

<210> 40
<211> 730
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 442401.2

<220>
<221> unsure
<222> 631
<223> a, t, c, g, or other

<400> 40
gcggcttccg ggatttggcg gtggcctttg ttggctgcag taagagctca gtctcttcac 60
caggggctcc cagtccttcc atctgggagg ccaaggcggc ttcgcgttct gagaatagac 120
agaacctctg ttactctgtg accggcaggc accgggagat ccgtagctca gacgccagga 180
catcccggaa gctgggaaat ggtgaatgtg ccagggactg ttgacattca gggatgtggc 240
catagaattc tctcgggagg agtgggaaca cctggactca gatcagaagc ttttatatgg 300
ggatgtgatg ttagagaact acggaaacct ggtctctctg ggtctcgctg tctctaagcc 360
ggacctgatc acctttttgg agcaaaggaa agagccctgg aatgtgaaga gtgcagagac 420
agtagccatc cagccagata tcttttctca tgatactcaa ggcctcttaa gaaagaagct 480
tatagaagca tcattccaaa aagtgatatt ggatggatat gggagctgtg gccctcagaa 540
tttaaactta aggaaagagt gggaaagtga gggcaaaata atcctatggt gaaaaaaaat 600
caacaagata atctctgcat gagaaaaagg nccaaagaca agtgaatttt ctgagggtga 660
tgaaagtgtt gcttcgaaaa ggggtgaagt ttatatgggt ctatttattt gtctaacatg 720
tacagttaag 730

<210> 41
<211> 575
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 444933.2

<400> 41
gggtccggcg ctccagaaca gaacgatccc tgaggctccc ttgctcgaac tgtgggactt 60
accctactat ggtccgagcc taccctattt cattatactc aagtaacgcc ccagaaattc 120
cagagaatct cacacaaaga ggttgagtct tgccgtggtg ccttcagggg aatgtcatcc 180
cgggctagaa gagctgcaaa aggctgtcag gcttctcaga actttgcttc tccagcagaa 240
taatcctgcg gaagactgag cagttcttgt gagtgtaaaa ccatggccca tgcattggtg 300

acgttcaggg atgtggctat agacttctct cagaaggaat gggagtgcct ggacactacc 360
cagaggaaat tgtacagaga tgtgatgttg gagaattata ataacttggt ctcactggga 420
tattctggct caaagccaga tgtgattacc ttactggagc aagggaaaga gccctgcgtg 480
gtggcgaggg atgtgacagg aagacagtgc cccggtttgt tatccaggca taagaccaag 540
aaattatctt cagaaaagga cattcatgaa atcag 575

<210> 42
<211> 734
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 481129.4

<400> 42
ggcgtgcgag actcggcggg cgctgttgag ggagtcgggc cgcgactgtg gtcgttttta 60
taccttcccg cgcggacgcc ggcgctgcca acggaagggc gggtagggcg gtgcgtgatt 120
aggttggcga agagacggag tttcgtcatg ttggccaggc ccatttgaga tctttgaaga 180
tatcctcaac gtgaggctct gctgccatga aggtgaagat taagtgctgg aacggcgtgg 240
ccacttggct ctgggtggcc aacgatgaga actgtggcat ctgcaggatg gcatttaacg 300
gatgctgccc tgactgcaag gtgcccggcg acgactgccc gctggtgtgg ggccagtgct 360
cccactgctt ccacatgcat tgcatcctca agtggctgca cgcacagcag gtgcagcagc 420
actgccccat gtgccgccag gaatggaagt tcaaggagtg aggcccgacc tggctctcgc 480
tggaggggca tcctgagact ccttcctcat gctggcgccg atggctgctg gggacagcgc 540
ccctgagctg caacaaggtg gaaacaaggg ctggagctgc gtttgttttg ccatcactat 600
gttgacactt ttatccaata agtgaaaact cattaaacta ctcaaatctt gctggaggcc 660
tctgggtgcc tgtgttctcg gcatatagat gtggtctcgg tgtgttttga tatgaaaact 720
ctaaatgaat aaac 734

<210> 43
<211> 1104
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 481999.1

<220>
<221> unsure
<222> 50, 201, 298, 905-906, 955, 961, 966, 979, 1001, 1021, 1028, 1032,
1074
<223> a, t, c, g, or other

<400> 43
ctgtcatcca tgctgatatg ctttcatttt gctgtgatag ggagagaatn cacatgtgct 60
tctcggaatt tgtgctgtca ctggagcatt tatttgagtg tctataatgt attcagggtt 120
gtgctaaaac ctgtcagtag agaacaataa gaagaagtca caatctagtt gtgctgagat 180
atattcagtt aggtgaaaga ngtgctcttc aagccatatg actagatgag atcaccggtg 240
aatatgtaca gacaggggag aagaaagtga ggaccatcaa gtgataggag accaaagngt 300
cttgtcaagg catgccaaga caaggtccct cattctccag tcaacatagt ggagaacgta 360
cttgcctgac gccagcatgt gtgtgatggt ttaggactca gtggctttag aggatgtggc 420
tgtgaacttc acccgagaag agtgggcttt gctgggtcct tgtcagaaga atctctacaa 480
agatgtgatg ctggagaact acatgaacct ggcctctgtg gaatgggaaa tacaacctag 540
aaccaaacgg tcatcacttc agcagggttt tttgaagaat caaatattca gtgggataca 600
aatgacaaga ggctacagtg gatggaaact ctgtgactgt aagaattgtg gagaggtctt 660
cagggaacag ttttgcctta agacacacat gagagttcag aatggaggga atacttctga 720
gggtaattgt tatggaaaag acaccctcag tgtgcacaag gaagcctcta ctggacagga 780
actttccaaa tttaatccat gtggaaaagt ctttactcta actccaggtc ttgctgtaca 840
tcttgaagtt ctcaatgcaa gacaacccta caaatgtaag gaatgtggaa aaggctttaa 900
gtatnntgca agccttgata atcatatggg aatccacact gatgagaaac tctgngaatt 960
ncaggnatat gggagagcng tcacagcttc ttcacaccta nagcagtgtg tagcagttca 1020
nacagganag anatccaaaa agactaagaa atgtgggaaa tccttcacta attnttctca 1080
actttatgca cctgtgaaaa ctca 1104

<210> 44

<211> 665
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 233814.1.j

<400> 44
ggagtgtgga agagccatcc agagtgagtg cctggggcag agtgagagca ggaccagcgg 60
gcaggtggcg gacgggggtc ggtggtggat ggcccagccc tggaaagccg aggtgaaggt 120
tacgagtccc tgaggatctg tgggtgtgag ccagccgcgt ggccctgggc aacggcgctg 180
ccggacagaa gtggcggttg ctgacgcctg gaaattcccc tgaaggtgga gcaccaccca 240
acccccctgg gtcccaccct ccctcaaggc ctcctccacc tccacctcca ccccgcctgg 300
cctggcgtcc acctctgcgg ctcctacctg ggtgcaatcg agttaaatgg ctgataagca 360
gatcagcctg ccagccaagc tcatcaatgg cggcatcgcg gggctgatcg gtgtcacctg 420
cgtgtttacc atcgacctgg ccaagacaag gctgcagaac cagcagaacg gccagcgcgt 480
gtacacgagc atgtccgact gcctcatcaa gaccgtccgc tccgagggct acttcggcat 540
gtaccgggga gctgctgtga acttgaccct cgtcaccccc gagaaggcca tcaagctggc 600
agccaacgac ttcttccgac atcagctctc taaggacggg cagaagcttg accctgcttt 660
aaaga 665

<210> 45
<211> 580
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 351376.4.j

<400> 45
agacactgct tgctgcggca gagacgccag aggtgcagct ccagcagcaa tggcagtgac 60
ggcgttggcg gcgcggacgt ggcttggcgt gtggggcgtg aggaccatgc aagcccgagg 120
cttcggctcg gatcagtccg agaatgtcga ccggggcgcg ggctccatcc gggaagccgg 180
tggggccttc ggaaagagag agcaggctga agaggaacga tatttccgac attacaggtt 240
atgctttgag atctctttgg ggtgaaggat tgaaattaaa ccctgagcca ccgtgtcctt 300
gtagagcaca gagtagagaa caactggcag ctttgaaaaa acaccatgaa gaagaaatcg 360
ttcatcataa gaaggagatt gagcgtctgc agaaagaaat tgagcgccat aagcagaaga 420
tcaaaatgct aaaacatgat gattaagtgc acaccgtgtg ccatagaatg gcacatgtca 480
ttgcccactt ctgtgtagac atggttctgg tttaactaat atttgtctgt gtgctactaa 540
cagattataa taaattgtca tcagtgaact gtgaaaaaaa 580

<210> 46
<211> 1935
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 338992.1

<220>
<221> unsure
<222> 52, 55, 83, 88, 96, 208, 511
<223> a, t, c, g, or other

<400> 46
tgaaagtact tttactccca agggtatctg atctgatgga gagaatgtat cnctngtacc 60
gatgcggata gaacgctcat tanttacngc acagtnttca taaattccaa cttccctagg 120
aattttgtct tttgaaggca tgttggaagt tggggcatta ggatataatc tcagtccatg 180
tactaagcct ctgcataact gaaagatnga gtcaataaaa ttcactctct catttttgt 240
ttttaatctc aagaattttt gccttgaagg aaaaatgtag tggagaattt gatgtttgaa 300
cgaatcccag tcagaagttc cagcctgcca ctgttctctg atgccatgcc agcaccaact 360
caactgtttt ttcctctcat ccgtaactgt gaactgagca ggatctatgg cactgcatgt 420
tactgccacc acaaacatct ctgttgttcc tcatcgtaca ttcctcagag tcgactgaga 480
tacacacctc atccagcata tgctaccttt ngtcaggcca aaggagaact ggtggcagta 540

cacccaagga aggagatatg cttccacacc acatacagtc cccctcacac ctccgcaagt 600
caatagcatc cttaaagcta atgaatacag tttcaaagtg ccagactttg acggcaaaaa 660
tgtcagttct atcctgtcga tttgacagca atcagctgcc tgcaaatgca cccattgagg 720
accggagaag tgcagcaacc tgcttgcaga ccagagggct ccttttgggg gtttttgatg 780
gccatgcagg ttgtgcttgt tcccaggcag tcagtgaaag actcttttat tatattgctg 840
tctctttgtt accccatgag actttgctag agattgaaaa tgcagtggag agcggccggg 900
cactgctacc cattctccag tggcacaagc accccaatga ttactttagt aaggaggcat 960
ccaaattgta ctttaacagc ttgaggactt actggcaaga gcttatagac ctcaacactg 1020
gtgagtcgac tgatattgat gttaaggagg ctctaattaa tgccttcaag aggcttgata 1080
atgacatctc cttggaggcg caagttggtg atcctaattc ttttctcaac tacctggtgc 1140
ttcgagtggc attttctgga gccactgctt gtgtggccca tgtggatggt gttgaccttc 1200
atgtgggcca atactggcga tagcagagcc atgctgggtg tgcaggaaga ggacggctca 1260
tggtcagcag tcacgctgtc taatgaccac aatgctcaaa atgaaagaga actagaacgg 1320
ctgaaattgg aacatccaaa gagtgaggcc aagagtgtcg tgaaacagga tcggctgctt 1380
ggcttgctga tgccatttag ggcatttgga gatgtaaagt tcaaatggag cattgacctt 1440
caaaagagag tgatagaatc tggcccagac cagttgaatg acaatgaata taccaagttt 1500
attcctccta attatcacac acctccttat ctcactgctg agccagaggt aacttaccac 1560
cgattaaggc cacaggataa gtttctggtg ttggctactg atgggttgtg ggagactatg 1620
cataggcagg atgtggttag gattgtgggt gagtacctaa ctggcatgca tcaccaacag 1680
ccaatagctg ttggtggcta caaggtgact ctgggacaga tgcatggcct tttaacagaa 1740
aggagaacca aaatgtcctc ggtatttgag gatcagaacg cagcaaccca tctcattcgc 1800
cacgctgtgg gcaacaacga gtttgggact gttgatcatg agcgcctctc taaaatgctt 1860
agtcttcctg aagagcttgc tcgaatgtac agagatgaca ttacaatcat tgtcgttcag 1920
ttcaattctc atgtt 1935

<210> 47
<211> 1709
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 200587.3.j

<400> 47
gcccggcagc agtggggcgg ggatggaggc ggccgtggcg ccggggaggg atgcgccggc 60
acccgcggcg agtcagccca gcggctgcgg gaaacacaac tcgccggaga ggaaagttta 120
tatggactat aatgcaacga ctcccctgga gccagaagtt atccaggcca tgaccaaggc 180
catgtgggaa gcctggggaa atcccagcag cccgtattca gcaggaagaa aggccaagga 240
tattataaat gcagctcggg aaagcctcgc gaagatgata gggggaaac ctcaagatat 300
aatcttcact tccgggggca ctgagtcaaa taatttagta atccattctg tggtgaaaca 360
tttccacgca aaccagacct caaagggaca cacaggtggg caccacagcc cagtgaaggg 420
ggccaagccc catttcatta cttcctcggt ggaacacgac tccatccggc tgcccctgga 480
gcacctggtg gaagaacaag tggcagcggt caccttgtc ccggtgtcca aggtgagcgg 540
gcaggcagag gtggacgaca tcctcgcggc agtccgcccg accacacgcc tcgtgaccat 600
catgctggcc aacaatgaga ctggcattgt catgcctgtc cctgaaatca gtcagcgcat 660
taaagccctg aaccaggaac gggtggcagc tgggctacct cccatcctcg tgcacacgga 720
tgctgcacag gccttgggga agcagcgcgt ggatgtggag gacctgggcg tggacttcct 780
tacaatcgtg gggcacaagt tttatggtcc caggattggc gcactttata tacgaggact 840
tggtgaattt acccctctct accctatgct atttggaggt ggacaagaac ggaatttcag 900
gccagggaca gagaacaccc caatgattgc tggccttggg aaggtgagcc ctggtgagct 960
tgaggagatg agttgattgc cttgtcatca gtgtccccag cggctctagt aggaatctgg 1020
ggctttaggg agatcatatt cgtgatctca taagcggact gacaagaaaa agcctggaat 1080
ctgatttatc aacctggaga gctgtacttt ttcagatctg tggattagtg gcacttgttt 1140
agcatttaaa agggaaaaac tccttgtctg tgtctcttct gacatgagaa acagtgatac 1200
tgttacagca gaggtgaagt gtcaagctgt aggttcacca ttcccaaagg cgtcggcagc 1260
ccccgcacca tcgacgttct gtccccgatt ggtctgtggg gtggccaatc cagcacctgc 1320
caccagccct ccactggggc ctgtcctgct tctgcgacag taaccaggct caggttttgc 1380
cgtgcgaagg cagaggtact ggttacacct cggcgtctgc aggctgctcc ctcactgtgc 1440
ggcctctggt caaaggccac caccttcaag caaccctcca tgaaccacag ctacattggt 1500
tcccccatgt cccctgctgt caccctctat ccctgacact gctttatttc ccctatctga 1560
tgtttctctg tagtgtgttc agatgtatgc cccaccttct gcttccccga gagcaggggc 1620
ttggccatgt cggcctcagt atccccagta cagagaaagg gcttagcgcc caggaggccg 1680
ggccttggga agagtatgaa tggattccc 1709

<210> 48
<211> 766

<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 246727.5.j

<400> 48
ggccgaggtg cgggtcgcct ccagaggtgc gtggtcgtgg cgcgagggat cctgaggctg 60
ctccagcagt gcgccgccgc cgtctcctgg ggcggcttgg gttagccggg aggtgggtca 120
gaggctcgcc ccgtgccctg cgtccgggcg tctccttagg gcgcgtcttc gggtccggcc 180
agcggtgctg aaaaagggag aaggttgggg gtagggagga aacaagatcc cagttcaata 240
gatttctccg cagatcctgt gccttcaaac cctacgagtc catactttaa aacaaaatga 300
agaaagtaag gcttaaggaa ctagagagtc gcctgcaaca agtggatgga tttgaaaagc 360
ccaagctact tctggaacag tatcctacca ggccgcacat tgcagcatgt atgctctata 420
caatccataa cacttatgat gacattgaaa ataaagtcgt tgcagatcta ggatgtggtt 480
gtggagtact tagcatcgga actgcaatgt taggagcagg ggacagatat ggcttttcta 540
aagactgctt tggaaatggc aagaacagca gtatattcct tacacaaatc ctcaactaga 600
gaacatgttc aaaagaaagc tgcagaatgg aaaatcaaga tagatattat agcagaactt 660
cgatatgacc tgccagcatc atacaagttt cacaaaaaga aatcagtgga cattgaagtg 720
gacctaattc ggttttcctt ttaaaagccc cgcaaacaaa agtcgt 766

<210> 49
<211> 2757
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 407087.3.j

<220>
<221> unsure
<222> 889, 917, 1054, 1098, 1122
<223> a, t, c, g, or other

<400> 49
gcgggagcaa gtgcggctgg agctgagctt cgtcaactca gacctgcaga tgctcaagga 60
agagctggag gggctgaaca tctcggtggg cgtctatcag aacacagagg aggcatttac 120
gattcccctg attcctcttg gcctgaagga aacgaaagac gtcgactttg cagtcgtcct 180
caaggatttt atcctggaac attacagtga agatggctat ttatatgaag atgaaattgc 240
agatcttatg gatctgagac aagcttgtcg gacgcctagc cgggatgagg ccggggtgga 300
actgctgatg acatacttca tccagctggg ctttgtcgag agtcgattct tcccgcccac 360
acggcagatg ggactcctgt tcacctggta tgactctctc accggggttc cggtcagcca 420
gcagaacctg ctgctggaga aggccagtgt cctgttcaac actggggccc tctacaccca 480
gattgggacc cggtgtgatc ggcagacgca ggctgggctg gagagtgcca tagatgcctt 540
tcagagagcc gcaggggttt taaattacct gaaagacaca tttacccata ctccaagtta 600
cgacatgagc cctgccatgc tcagcgtgct cgtcaaaatg atgcttgcac aagcccaaga 660
aagcgtgttt gagaaaatca gccttcctgg gatccggaat gaattcttca tgctggtgaa 720
ggtggctcag gaggctgcta aggtgggaga ggtctaccaa cagctacacg cagccatgag 780
ccaggcgccg gtgaaagata acatccccta ctcctgggcc agcttagcct gcgtgaaggc 840
ccaccactac gcggccctgg cccactactt cactgccatc ctcctcatng accaccaggt 900
gaagccaggc acggatntgg accaccagga gaagtgcctg tcccagctct acgaccacat 960
gccagagggg ctgacaccct tggccacact gaagaatgat cagcagcgcc gacagctggg 1020
gaagtcccac ttgcgcagag ccatggctca tcangaggag tcggtgcggg aggcgagcct 1080
ctgcaagaag ctccgganat tgaggtgcta cagaaggtgt gngtgccgca caggaacgct 1140
cccggctcac gtacgcccag caccaggagg aggatgacct gctgaacctg atcgacgccc 1200
ccagtgttgt tgctaaaact gagcaagagg ttgacattat attgccccat tctccaagct 1260
gacagtcacg gacttcttcc agaagctggg ccccttatct gtgttttcgg ctaacaagcg 1320
gtggacgcct cctcgaagca tccgcttcac tgcagaagaa ggggacttgg ggttcacctt 1380
gagagggaac gcccccgttc aggttcactt cctggatcct tactgctctg cctcggtggc 1440
aggagcccgg gaaggagatt atattgtctc cattcagctt gtggattgta agtggctgac 1500
gctgagtgag gttatgaagc tgctgaagag ctttggcgag gacgagattc gagatgaaag 1560
tcgtgagcct cctggactcc acatcatcca tgcataataa gagtgccaca tactccgtgg 1620
gaatgcagaa aacgtactcc atgatctgct tagccattga tgatgacgac aaaactgata 1680
aaaccaagaa aatctccaag aagctttcct tcctgagttg gggcaccaac aagaacagac 1740
agaagtcagc ccagcacctt gtgcctccca tcggtcgggg ctgcacggcc tcaggtcaag 1800

aagaagctgc cctccccttt cagccttctc aactcagaca gttcttggta ctaatgtgag 1860
gaaacaaaca tgttcaggcc ctgaacattt ccggtgctga ctcggcctta aacgtttgtg 1920
ccataatgga aaatatctat ctatctgttc tcaaatcctg tttttctcat agtgtaaact 1980
cacatttgat gtgtttttat gaaggaaagt aaccaagaaa cctctaggaa ttagtgaaaa 2040
aagaactttt ttgaggtgtg ttactatact gctgtaagtt atttattata taaagtattg 2100
taaatagaat agtgttgaag atatgaaata tggctatttt taatggtgac aattatgact 2160
tttagtcact attaaattgg ggttacctat atcagtacaa tttgtagttg tttccaggtt 2220
tggctaataa tcattcctta acctagaatt cagatgatcc tggaattaag gcaggtcaga 2280
ggactgtaat gatagaatta aattagtgtc actaaaaact gtcccaaagt gctgcttcct 2340
aataggaatt cattaaccta aaacaagatg ttactattat atcgatagac tatgaatgct 2400
atttctagaa aaagtctagt gccaaatttg tcttattaaa taaaaacaat gtaggagcag 2460
cttttcttct agtttgatgt catttaagaa ttactaacac agtggcagtg ttagatgaag 2520
atgctgtcta caaggtagat aatatactgt ttgatactca aaacattttt cattttgttt 2580
aaagtagaag ttacataatt ctatatttta agtcttgggt aaaaaagtag ttttacattt 2640
tataaagtaa agatgtaaat gattcaggtt taaagctcta tttgacttcc ttttttgtt 2700
tgagatagcg tcttgctgtg ttgcccaggc tgggagtgca gtgggtgtga tctcgag 2757

<210> 50
<211> 558
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 441779.1.j

<400> 50
atggaggagc tatggacgtc gcaatgcacg cgtacgtaaa gctcggaatt cggctcgagg 60
actcaggaag caatcatggt gctctctgca gctgacaaaa ccaacatcaa gaactgctgg 120
gggaagattg gtggccatgg tggtgaatat ggcgaggagg ccctacagag gatgttcgct 180
gccttcccca ccaccaagac ctacttctct cacattgatg taagccccgg ctctgcccag 240
gtcaaggctc acggcaagaa ggttgctgat gccctggcca aagctgcaga ccacgtcgaa 300
gacctgcctg gtgccctgtc cactctgagc gacctgcatg cccacaaact gcgtgtggat 360
cctgtcaact tcaagttcct gagccactgc ctgctggtga ccttggcttg ccaccaccct 420
ggggatttca cacctgccat gcacgcctct ctggacaaat tccttgcctc tgtgagcacc 480
gtgctgacct ccaagtaccg ttaagccacc tcctgtcggg cttgccttct gaccaggccc 540
ttcttccgtc ccctgaac 558

<210> 51
<211> 905
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 206603.1

<220>
<221> unsure
<222> 374
<223> a, t, c, g, or other

<400> 51
gggccgcggc cggcagaagg gctgttagga gggaccacgc gccgggggcc gcgatctctg 60
gcaggggggcg gtgtgccagc ggagcaccat gcacataggc gcccagcgcc ccgactaccc 120
ctcccgagga aaagaggccg gggccgcgct ggggcggcgg agagcatgag ggaggccggg 180
gggcggctcg gcttggagcg ctgctaggga gcggtgcgcg ccgcacaccc gcctgggcgc 240
ggcggagggc ggggagccgg gcaggtcgcg cctgcgggcg gcagccgacc gccgggagct 300
gttctgattt ccgacgcgca cgctaggggc ccggagcagc ccccggcccc ggcgcgccgc 360
cgacatgggc aacngcaggg agcattggat tcgcagcaga ccgatttcag ggcgcacaac 420
cgtgcctttg aagctgccga tgccagagcc aggtgaactg gaggagcgat ttgccatcgt 480
gctgaatgct atgaacctac ctcctgacaa agccaggtta ctgcggcagt atgataatga 540
gaaaaaatgg gaactgattt gtgatcagga acgattccag gtgaagaatc ctccccatac 600
atacattcaa aagctcaaag gctatctgga tccagctgta accaggaaga aattcagacg 660
gcgtgttcaa gaatctacac aagtgctaag agaactggaa atttctttga gaactaacca 720
cattggatgg gtcagagaat ttctgaatga agaaaacaaa ggtcttgatg ttctagtgga 780
atatctctca tttgcacagt acgcgagaac ttttgacttt gaaagtgtgg agagtactgt 840

ggagagctcg gtggacaaat caaagccctg gagtaggtcc atcgaggacc tgcacagagg 900
gagca 905

<210> 52
<211> 2160
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<223> Incyte ID No: 435694.2

<220>
<221> unsure
<222> 484, 488, 491, 494
<223> a, t, c, g, or other

<400> 52
cgttcttttc ctttctcctt gaaagagtga atgcatttgg tcgcagggct aaagaggagg 60
atgctatact tttctaaatg gcaagagatg gggagagaag gggattaaga gttgacccgc 120
aacctcccgg ggattctttg ttcttaccag atctcttggc cactccccta ttctgaagtc 180
gtcttggctc tcttgactgc tcccctattc tgaagtcgtc ttggctctct tgactgctcc 240
cctattctga agtcgtcttg gctctcctga ctacactatt tcaaggaatg atcaccaaga 300
cacacaaagt agaccttggg ctcccagaga agaaaaagaa gaagaaagtg gtcaaagaac 360
cagagactcg atactcagtt ttaaacaatg atgattactt tgctgatgtt tctcctttaa 420
gagctacatc cccctctaag agtgtggccc atgggcaggc acctgagatg cctctagtga 480
aganaaanaa naanaaaaag aagggtgtca gcacccttg cgaggagcat gtagaacctg 540
agaccacgct gcctgctaga cggacagaga agtcacccag cctcaggaag caggtgtttg 600
gccacttgga gttcctcagt ggggaaaaga aaaataagaa gtcacctcta gccatgtccc 660
atgcctctgg ggtgaaaacc tccccagacc ctagacaggg tgaggaggaa accagagttg 720
gcaagaagct caaaaaacac aagaaggaaa aaaagggggc ccaggacccc acagccttct 780
cggtccagga cccttggttc tgtgaggcca gggaggccag ggatgttggg gacacttgct 840
cagtggggaa gaaggatgag gaacaggcag ccttggggca gaaacggaag cggaagagcc 900
ccagagaaca caatgggaag gtgaagaaga aaaaaaaaat ccaccaggag ggagatgccc 960
tcccaggcca ctccaagccc tccaggtcca tggagagcag ccctaggaaa ggaagtaaaa 1020
agaagccagt caaagttgag gctccggaat acatccccat aagtgatgac cctaaggcct 1080
ccgcaaagaa aaagatgaag tccaaaaaga aggtagagca gccagtcatc gaggagccag 1140
ctctgaaaag gaagaaaaag aagaagagga aagagagtgg ggtagcagga gacccttgga 1200
aggaggaaac agacacggac ttagaggtgg tgttggaaaa aaaaggcaac atggatgagg 1260
cgcacataga ccaggtgagg cgaaaggcct tgcaagaaga gatcgatcgc gagtcaggca 1320
aaacggaagc ttctgaaacc aggaagtgga cgggaaccca gtttggccag tgggatactg 1380
ctggttttga gaacgaggac caaaaactga aatttctcag acttatgggt ggcttcaaaa 1440
acctgtcccc ttcgttcagc cgccccgcca gcacgattgc aaggcccaac atggccctcg 1500
gcaagaaggc ggctgacagc ctgcagcaga atctgcagcg ggactacgac cgggccatga 1560
gctggaagta cagccgggga gccggcctcg gcttctccac cgcccccaac aagatctttt 1620
acattgacag gaacgcttcc aagtcagtca agctggaaga ttaaactcta gagttttgtc 1680
cccccaaaac tgccacaatt gctttgatta ttccatttat gctggagatt acaaattttt 1740
tttgtgaaaa aatcagatct tggtgaggac ctcgagcagt aagatataaa taactcccat 1800
aagcttagcg ttccagtaat ggaacactag gcataaatgg tttattcagt tgtgcaaatg 1860
aaagccatct gacagttggc tcacattgaa cacctgtgga gattaaggac gaggacaact 1920
atattgatgg gcttggatga actggggcag ggcagctcat atttcgggag ccaggagaac 1980
gagtgagtgc taaaacctcc tgttttctgt gttaaacatt ccgtccctgt ttgagacatc 2040
agtatgtaca gttaactttt gttgagtgtt tagcaggtac tagggacata ctagtgtttt 2100
ccttaatgta tttaatcttc ataattatga aatgggtgct attattagcc ccatcttata 2160